An evidence-centred approach to Reverse Engineering: Comparative analysis of IELTS and TOEFL iBT reading sections

Thesis submitted for the degree of Doctor of Philosophy at the University of Leicester

By

Nathaniel Owen

School of Education

University of Leicester

2016

Acknowledgements

This PhD would not have been possible without the invaluable assistance of a number of people. In particular, I would like to acknowledge the help and support of the following people:

My supervisor, Professor Glenn Fulcher, a friend and mentor who always helped me to stay grounded and was a source of constant inspiration.

The love of my life, Haiyan Xu, the rock upon which this thesis is built. Writing a thesis can be a lonely experience. I would not have been able to complete it without her.

My family, a source of strength in my life. They may be scattered to the corners of the Earth, but are always willing to find time for me in any hour of need.

Members of the academic community. I have had the opportunity to attend a number of conferences and seminars related to language testing, and have either attended or presented at the following: the University of Bedfordshire Centre for Research in English Language Learning and Assessment (CRELLA) research seminars, Language Testing Forum (LTF) in Nottingham, Bristol, Southampton and Oxford and the Language Testing Research Colloquium in Princeton, Amsterdam and Toronto.

The University of Leicester, who employed me as a Research Assistant, thus paving the way for me to pursue a full-time PhD.

Educational Testing service (ETS), who provided the small doctoral grant for this study in 2013. This was invaluable for completing the study and was a huge boost to my confidence to know that such a prestigious company believed in my research.

During moments of indecision, writers' block, or lacking motivation and requiring a change of scenery, a common occurrence was to retreat to a café for a latte and a muffin. I believe I have now gained a greater understanding of a moment in George Bernard Shaw's play *Man and Superman*, in which the main character, John Tanner, laments that "one never tires of muffins, but it is impossible to find inspiration in them".

Contents

List of tables and figures

List of figures

An evidence-centred approach to Reverse Engineering: Comparative analysis of IELTS and TOEFL iBT reading sections

Chapter 1. Introduction

1.1. The prevalence of high-stakes tests of English as a foreign language and the problem of test comparability

Universities in English-speaking countries require prospective students from those countries which do not speak English as a first language to provide evidence of their English language proficiency prior to the commencement of academic study. Evidence regarding this claim is presented to the academic institution in the form of a test score. This score is then used as part of the admissions process to determine the relative success or failure of each applicant. Until recently, students wishing to undertake a programme of study in an English-speaking country would take whichever test the institution to which they were applying required: this was invariably split along geographical lines – US institutions requiring a score from the *Test of English as a Foreign Language Internet-based Test*™ (TOEFL iBT®), with those in the UK requiring a score from the *International English Language Testing System*™ (IELTS®).

More recently, however, the respective 'spheres of influence' of the two educational blocs have become increasingly blurred as institutions have become more flexible in deciding which scores to accept from which testing companies. Therefore, potential applicants are faced with a choice of which test to take. The importance of the outcome of taking a test of English for entry to higher education means this is an extremely high-stakes decision. In the past, test takers took whichever test was approved by their institution of choice. Now, they are faced with a choice and a paucity of information with which to make this choice. Test takers trust that they will have a fair chance of success whichever test they choose. This trust is based on two principles. First, that the acceptable levels of performance in the tests are set at

equivalent levels, such that performance in either test is equally likely to result in a successful outcome by an individual. Secondly, that the tests have an identical, or at least highly similar conception of the construct of 'English for academic purposes' (EAP). The first of these issues has received some attention in the literature. The second has received almost none.

1.2. The limited quality of publicly available information on test construct definitions

Comparability studies to date have focused on comparing scores obtained from cohorts of test takers that have completed both tests, or by comparing verbal descriptions of abilities in scoring rubrics with CEFR verbal descriptors. As testing companies do not indicate a pass/fail grade, individual institutions have to determine levels of equivalence between tests. Once an acceptable level of performance is determined by an institution, this is then applied to the scoring rubrics of multiple tests and a de facto level of equivalence between tests is established. However, these institutional cut scores are rarely based on independent research: in fact, most 'equivalent cut scores' are decided on the basis of existing correspondence tables that have little or no empirical basis. These are often posted on the internet without referencing source studies (e.g. http://secure.vec.bc.ca/toefl-equivalency-table.cfm). As a result, organisations which create these tests have produced research that compares the tests directly based on their scoring rubrics and score reports of test takers (ETS 2007, 2010) and via proxy measures such as the *Common European Framework of Reference* (CEFR). This research is highly technical and focuses on statistical issues of comparing scoring rubrics, while ignoring the extent to which the tests being compared have similar understandings of the construct.

There is little publicly available information regarding test construct definitions that is easily accessible to stakeholders. Therefore, there is little by way of information for test takers to readily compare different tests to determine what they understand by 'reading for academic purposes', and how this links to test design and content. Different tests may claim to measure the same construct, but will be designed and

delivered in very different ways. Test takers therefore have no information regarding whether the different design is simply an alternative means of accessing the same construct, or whether there is something fundamentally different in how the construct has been conceived by the test developers.

Information regarding how the construct has been operationalised by the developers is often included in the *test specification*. Test specifications are "generative blueprints"; documents which inform test design and content, and may be used iteratively in the creation of carefully-designed test items (Fulcher and Davidson, 2007: 377). However, test specifications are proprietary information; they are confidential documents used for internal purposes and not written for public consumption. Greater access to test specification documents produced by developers would be beneficial for stakeholders. Test takers would have greater information about the nature of the construct and how it had been operationalised. Universities would have more information to make decisions regarding which tests are most suitable for their admissions policies and there would be far less scope for test misuse. The limited information regarding test developers' conception of the construct of interest is a longstanding problem which this thesis aims to address.

1.3. Reverse engineering as a solution to the problem of limited information regarding test construct definitions

Reverse engineering (RE) offers one solution to this problem. RE is a means of creating a test specification where one is unavailable. Either a test was developed without an initial specification, perhaps due to institutional pressures such as a lack of resources, or the specification is unavailable because it is regarded as proprietary information, as previously stated. RE is the post-hoc creation of a specification, created from representative examples of that test (Davidson and Lynch, 2002: 41). The impetus for this thesis is the desire to develop and test an RE framework that can be consistently applied to English language tests for the purposes of uncovering information about test developers' understanding of the construct embedded in their tests. A systematic approach to test analysis can result in the production of a representative specification.

Various stakeholders maintain an interest in specification documents and it is the position of this thesis that RE can support ethical test preparation by more clearly linking the construct to test design. Having a clear definition of the construct and knowing what elements of the construct are associated with particular items can assist teachers and publishers to create similar tasks that target those elements, and avoid preparation that focuses on construct-irrelevant strategic management of those tasks. This is good pedagogical practice, carefully documented, and provides the basis of assessment for learning.

1.4. Developing a theoretical understanding of reverse engineering

The concept of RE has received sparse attention in the literature, despite the proposed benefits. Davidson and Lynch (2002) are credited with creating the concept, although go no further than offering a definition. Fulcher and Davidson (2007: 57-58) have given the concept most attention to date, providing a general 'typology', although they provide no information about how to develop a principled research project, or what types of methodologies are most suitable.

However, RE has received attention in other academic disciplines. In computer science, the concept has been carefully developed over many years (Chikofsky and Cross, 1990; Eilam, 2005). The purpose of RE in computer science is to gain a design-level understanding of a program or system for maintenance, product enhancement or replacement (Chikofsky and Cross, 1990: 14). The concept has received detailed theoretical development. It is possible that the concept has not received much attention in language testing due to negative connotations of the term 'reverse engineering', which is immediately associated with industrial espionage and theft of intellectual property. However, RE as a procedure is neither fundamentally good nor bad, it is the purpose to which it is put that should be judged. The thesis uses the theoretical development of RE in computer science to develop the concept of RE and identify what can be learned that is of benefit to language testing.

1.5. Developing a comprehensive framework of reverse engineering using evidence-centred design (ECD)

The approach to conceptualising RE in this thesis was inspired by Fulcher and Davidson's (2009) use of architectural principles to reconceptualise the process of 'retrofitting' a test (i.e. changing a test to meet a new purpose). Mislevy et al (2003) and Mislevy and Riconscente (2005) have shown how architectural test layers contribute to test development in theory and practice. They term this process 'evidence-centred design' (ECD). The thesis uses the principles of 'evidence-centred design' (ECD) to further develop the RE framework. ECD is a formalised test design process which identifies different areas for which evidence must be provided in order to create a test validity argument (Mislevy et al, (2003; Mislevy and Riconscente, 2005). The part of ECD which was identified as most relevant in this thesis is called the *conceptual assessment framework* (CAF), containing student, evidence and task models. The student model (SM) describes the construct in terms of what is being measured (knowledge, skills and abilities). The evidence model (EM) details the evidence required to make inferences from observable variables (work products) to the construct of interest. The task models describe the material that is presented to the test takers and what the expected work product will look like. These models are integrated into the RE framework developed from the computer science literature to produce a framework of RE suitable for a research agenda of analysing language tests and identifying important aspects of the construct that are embedded within them.

As the framework is built upon ECD, so it shares the conception of validity. Validity in ECD is based on an interpretive epistemology. Specifically, the concept of validity is understood as a process of argument (Kane, 2006). This argument adopts a network of inferences, such that each piece of data gathered fits into a logical argument, which is used to link data to claims made on the basis of that data. Within this thesis, statements made about the construct embedded in a test become the claims. The RE framework therefore assists reverse engineers to identify their research aims, focus of investigation and what claims they can make about their test based on the data that they have collected.

1.6. Demonstrating the efficacy of the RE framework through comparative analysis of two tests of English as a foreign language

Having developed an evidence-centred framework of RE, the thesis seeks to demonstrate the utility of this approach to RE with a practical example. The research agenda selected to assess the framework is a comparative analysis of the reading sections of the *International English Language Testing System* (IELTS) and the *Test of English as a Foreign Language internet-based test* (TOEFL iBT). These two tests are used for high-stakes purposes, namely to make decisions about prospective applicants to universities in which English is the language of instruction. Test development of both IELTS and TOEFL has occurred through iterative cycles of feedback and reconsideration involving detailed analysis of sets of tasks in an attempt to elucidate clear design principles and standards for each task type. These processes result in a stable and well-developed test specifications being used internally for the production of equivalent forms of these tests. These tests are therefore suitable instruments for the development of an RE framework.

These tests are ostensibly used for the same purposes and are both accepted as evidence of English language proficiency by educational institutions attracting large numbers of international applicants. However, the question of the comparability of these two tests remains critically understudied. Thus, there are two key dimensions to this thesis: the development and testing of an emerging RE framework, and the comparability of IELTS and TOEFL in terms of how they operationalise the construct of English for academic purposes (EAP).

1.7. Limiting the scope of the study to the reading components of IELTS and TOEFL

A comprehensive RE research project encompassing a totality of reading, speaking, writing and listening components of two tests is beyond the scope of even a PhD thesis. Therefore, I decided that the research would be limited to the reading

components of IELTS and TOEFL iBT. Reading is a key element of university study. The ability to critically engage with source material on higher education courses is key to academic success. This is more difficult for individuals who are taking courses in a second or even a third language. The importance of reading is evident in the central role that this skill occupies in the IELTS and TOEFL tests. Moore et al (2007) note that despite the significant changes that have been undertaken in major language tests over the last 30 years, including the development of an 'integrated skills approach' by ETS, both IELTS and TOEFL tests have retained dedicated sections for reading. As high-stakes tests require a significant amount of information from participants to be able to make reliable and valid claims about test takers' reading proficiency, the reading sections of both tests are lengthy (each contains three texts) and contain a large number of items and item types.

Hawkey (2006) has called for further investigation into the validity of the IELTS reading test on the basis that he identified discrepancies between reading as it exists in higher education in the United Kingdom and reading as it is measured in the IELTS test. As the TOEFL test is primarily used to make decisions about prospective students in the United States, this warrants investigation into whether academic reading in the US is aligned to the conception of academic reading in the TOEFL. This potential discrepancy provides the rationale for an in-depth comparative study of the reading components of IELTS and TOEFL iBT which this study will fulfil. A further justification for focusing on reading is the opaque nature of reading item selection and design, which will be clarified through RE. Reading items also provide prospects for rich analysis as test takers are required to mediate between questions and input texts. This interface provides a rich space in which to investigate the test developers' conception of reading for academic purposes.

1.8. Selecting an appropriate methodology for studying the reading components of IELTS and TOEFL

The research agenda for the RE is comparative analysis of IELTS and TOEFL, in order to compare their respective conceptions of the construct of reading for academic

purposes, to identify areas of congruence and difference. As these tests are frequently compared (ETS 2007, 2010), this thesis will present evidence of their respective construct definitions, to determine the extent to which a comparison of their score rubrics is justified. Specific research questions were developed from the constitutive student, evidence and task models from the RE framework. The student model refers to the construct that is represented in the test as a whole. The task model attempts to determine whether specific aspects of the construct are associated with particular items. The evidence model specifically identifies the observable evidence that is used to link data to claims about the construct. The study therefore required a methodology that could provide sufficient data for analysis so that claims could clearly be made about the respective constructs of the two tests.

The study adopted stimulated-recall interviews (SRI) as the main research methodology. SRI are a specific type of interview in which participants verbalise their thought processes in relation to a particular task, prompted with stimuli from their task performance. In this thesis, the stimuli consist of participants' engagement with the test items, in the form of responses, annotations, and highlighting of key words/terms. Additional video evidence provided data of when and how engagement occurred. Interview verbalisations are then analysed in one of two ways. They can be analysed from a grounded perspective, in which claims made about their performance emerge from the verbal data to form an understanding of task completion. Alternatively, they can be analysed from a pre-existing framework which was designed or adapted from existing literature. This study adopts both approaches. Observable actions taken by the participants are interpreted as metacognitive *strategies*. These are deliberate and goal-oriented actions which the test taker is aware of (and can report using) and considers pertinent to a specific task (Phakiti, 2003: 29). A strategy taxonomy emerged organically out of the data collected. Participants' verbalisations were analysed using a research instrument devised from a pre-existing model of reading. Participants' explanations were analysed for evidence of the *cognitive processing* they underwent to answer each item. Cognitive processing refers to the mental procedures that occurs when test takers perform a task, representing the test

taker's ability to engage with material presented to him or her with the linguistic resources available (language *use*).

From engaging with the literature regarding reading in English as a foreign language, this thesis adopted Khalifa and Weir's (2009) model of reading to analyse participant verbalisations. The model presents a cognitive hierarchy of engagement, from identifying individual orthographic forms of a word through to forming mental models of a text or combining mental models of two or more texts. This componential model was adopted for two important reasons. First, the model was amenable to transformation into a research coding instrument. Second, as the model is hierarchical, each level subsumes the previous levels. Therefore, as one progresses up the model, the processing load is thought to be greater, and therefore more difficult. This model is extremely useful from a comparative perspective, as tests can be compared directly in terms of their emphasis on lower or higher levels of processing.

1.9. Outcomes of the research: comparative claims regarding IELTS and TOEFL

The outcomes of the research will be presented in the form of cognitive test specifications for IELTS and TOEFL. This data is then amalgamated for comparative purposes to make comparative claims about the two tests. Observed strategies were described without reference to either test in order to create a generic strategic coding matrix applicable to a number of tests and research agendas. The video approach to the stimulated-recall methodology resulted in the identification of 1276 individual moments of participant decision-making for the six participants.

The data revealed subtle distinctions between the tests, some of which were due to test design and some to test content. Eliminating options was a more frequently observed action for TOEFL than IELTS, for example. This is likely the result of method effect due to the greater number of multiple-choice items in TOEFL, leading to a more consistent option-elimination strategy. Participants were also more likely to read the questions before proceeding to the text in TOEFL than in IELTS, reflecting greater

complexity in TOEFL item stems. Participants also underlined a much greater proportion of noun phrases in the TOEFL test compared to IELTS, suggesting that the TOEFL test is more likely to prioritise conceptual understanding, with IELTS more likely to highlight verb and adjective phrases.

Khalifa and Weir's (2009) model was applied to participant verbalisations. Each of the levels was given a code and verbalisations were analysed for evidence of the highest level of processing that could be inferred on the basis of participants' explanation of their actions and responses to specific items. During the analysis of the data, several of the categories were divided as it became apparent that the data was sufficiently finely-grained to allow for a greater number of codes. Neither test recorded any instances of the highest level of processing (creating inter-textual representation), which suggests that both tests have an incomplete definition of the construct of reading for academic purposes. Forming meaning across texts is a crucial skill in academic reading, which is ignored by both test developers. The data also revealed that TOEFL items are designed in specific ways to target particular cognitive processes. In contrast, the IELTS test has a range of item types, each of which is designed to gather evidence about a range of cognitive processes. Overall, the data suggests that the two tests actually contain a very similar conception of the domain of interest, as defined by the model of reading by Khalifa and Weir. Both tests require both higher and lower level processing to complete successfully.

1.10. The efficacy of the RE framework for addressing the research aims

Ideally, the best means of evaluating the RE framework would be to compare the outcome specifications directly to the original documents. However, this is not possible, as these are proprietary documents which are regarded as commercially sensitive. Therefore, the claims made about the constructs are evaluated in terms of the strength of association between data and claims. This association is predicated on ECD, a framework for linking claims specifically to data via Toulmin's (2003) model of argumentation. RE, as conceived in the present study, is subject to the same challenges associated with an evidence-based validation project. Evidence-centred

design, with its modular approach to test development and argument-based approach, fitted the agenda perfectly. The requirement of the evidence model in the RE framework drove methodological innovation which otherwise would not have occurred. This methodological innovation led to stronger claims being made about the nature of the construct of reading for academic purposes embedded in IELTS and TOEFL.

1.11. Outline of the thesis

The thesis contains five major chapters (the introduction, the literature review, methodology, findings and discussion and the conclusion). The literature review describes the theoretical development of the concept of RE through to a framework in which the analysis will take place. As the thesis is based on the study of reading, the literature review also includes an appraisal of the current thinking related to reading in a second language. The new RE framework and insight gained from the review of L2 reading literature resulted in the formulation of four guiding research questions, outlined at the end of the literature review.

The methodology opens with a short section containing the key terms and definitions gleaned from L2 reading literature that will inform the study. It then progresses to outline the methodology by which the research questions will be addressed (stimulated-recall interviews), and the philosophical implications of adopting this methodology. Practicalities of the interviews, the nature of the stimuli and the participants are all outlined. Three initial studies and one pilot study were conducted to refine and evaluate the methodology. The final methodological design is presented in relation to each of the four main research questions. The chapter includes an explanation of how an objective procedure was conducted to select specific versions of the two tests for the main study.

The findings and discussion chapter is divided into sections that relate to each of the four research questions. The first part details the observable evidence gathered from the video-stimulus, which is transformed into a strategy taxonomy. Similarities and

differences between IELTS and TOEFL are discussed in terms of how participants engage with the tests. Section 2 details the findings of the stimulated-recall interviews in terms of the cognitive processes that could be identified in the interviews. Video-stimuli were presented to participants to enable them to provide feedback on how they completed individual items. Video was also used by the researcher in conjunction with the coding framework developed from Khalifa and Weir's (2009) cognitive model of reading to code participant verbalisations. The third section discusses similarities and differences that emerged between the two tests and whether the findings reveal differences in the understanding of the construct of interest. The fourth section presents the individual *task models.* This section discusses findings related to individual items types in the study. Information is presented in both test specification format and discussion of the cognitive processes relevant to individual item types.

The thesis concludes by summarising the contribution of the thesis to our understanding of L2 reading, how a renewed concept of RE has contributed to the main findings, how the adoption of an RE framework led to methodological innovation, and a consideration of how the strengths and limitations of the study may inform future studies.

Chapter 2. Literature Review

2.1. Introduction

The literature review focuses on two main areas. First, the thesis is concerned with the development of a framework of RE. Literature that is relevant to this goal is explored. Second, as the thesis proposes to compare the reading components of IELTS and TOEFL, current research into the testing of second language reading is explored, with particular emphasis on what the test developers (Cambridge ESOL and ETS respectively) claim about their tests.

The study was motivated by questioning the comparability of IELTS and TOEFL. The chapter therefore starts by identifying and problematizing existing methods of test comparability. The review argues that to date, these studies have predominantly been too narrowly-focused on the technical details of *score* comparability, rather than focusing on how different conceptions of the construct of interest have resulted in different instrument designs and content. This understanding of test comparability remains critically under-explored. The study seeks to identify what differences in the construct definition of reading embedded in IELTS and TOEFL exist.

RE is proposed as a solution to this problem. The concept of RE emerged from systems engineering, and has been used in language testing literature as a means of identifying test design principles found in test specification documents where none were previously available. However, the language testing literature reveals that the concept suffers from under-development. To base a comparability study on RE, a detailed framework must be developed. As the language testing literature is insufficient for developing a comprehensive RE framework, the review considers what lessons can be learned from computer science, where the concept is much more carefully theorised. The skeleton of a RE model is taken forward from this literature. Additionally, the literature review considers how existing test development literature may provide the internal contents of this skeleton to form the comprehensive framework. The

literature review identifies relevant components of evidence-centred design that will provide the content of the framework.

As the study focuses on reading, the second part of the literature review considers how the testing of L2 reading is understood in order for the study to successfully identify constituent parts of the construct embedded in IELTS and TOEFL. This section focuses on understanding what defines L2 reading, how contemporary theory in the literature can be used, and how L2 reading is investigated. Specifically, this section examines recent literature which presents a cognitive perspective of reading. The section also explores how reading is understood by the test developers for IELTS and TOEFL. This section also considers how cognitive understandings of second language reading is studied, identifying and examining the strengths and weaknesses of verbal protocol analysis. The literature review concludes by integrating the understanding of a cognitive view of reading with the newly-developed RE framework. This results in the development of four critical research questions, which will guide the data collection and analysis.

2.2. Current approaches to test comparability research

Technical procedures of linking or comparing tests have received significant attention in the literature. The aim of linking tests is to be able to use scores on different scales interchangeably, so that decisions can be made about test takers who have taken different, but supposedly 'equivalent' tests. Studies may be concerned with direct linking of different tests that are reported on different scales, or researchers may be interested in regression-based prediction of performance on one test on the basis of another (Dorans and Walker, 2007: 179). However, such studies are usually politically or economically driven such as linking a test to an externally produced framework such as the *Common European Framework of Reference* for languages (CEFR), or may be necessary on the basis of comparing new versions of a test to older ones, as in the case of the paper-based TOEFL (PBT) and the TOEFL internet-based test (iBT) (Eignor, 2007: 152-153).

Holland and Dorans (2006) identify three principal measures of linking scores; predicting, scale aligning and equating. Predicting requires a cohort to have taken two tests. Performance across tests is then used to predict what score a new test taker would receive in test A based on their performance in test B. The oldest method of scale aligning is the equipercentile method (Kelly, 1923), in which score distributions for two tests are divided into 100. Percentiles are then compared across the two tests to identify equivalent scores. More recently, scale aligning converted test scores into standardised (z) scores in order to compare individuals in relation to their respective sample distributions. A score which was one standard deviation above the mean for a test was then compared to a score which was also one standard deviation above the mean for the second test. The type of scaling undertaken varies depending on the congruence of *reliability*, *difficulty* and *ability* of the test-taking population. *Equating* is the most comprehensive form of score linking and is undertaken when researchers are confident that each of these three variables is congruent (Holland, op cit.: 20). Item-response theory (IRT), based on the Rasch model (Rasch, 1960) is used for this purpose, as it can provide data on test difficulty and test taker ability in addition to reliability statistics. However, in terms of comparing tests, Mislevy (1992: 14. Emphasis added) provides a warning that "the degree to which linking can succeed, and the nature of machinery required to carry it out depend on the *matchups between the purposes for which the assessments were constructed and the aspects of competence they were designed to reveal*". Here, Mislevy endorses a research agenda which places construct definition at the heart of test comparability.

To date, there has only been one large-scale study which has attempted to directly compare international English-language admissions tests which has encompassed a focus on construct definition. Bachman et al (1995) investigated the comparability of the *Cambridge First Certificate in English*™ (FCE®), *Certificate of Proficiency in English*™ (CPE®) papers, and the paper-based TOEFL. Studies that compare the TOEFL iBT and IELTS (ETS, 2010) Pearson, TOEFL, IELTS and the CEFR (Pearson, 2009) and TOEFL and the CEFR (ETS, 2007) have specifically been concerned with estimates of *score* concordance, as outlined above. Each of these studies will be briefly addressed in terms of what they offer to the current research agenda.

Bachman et al (1995) investigated the comparability of the *Cambridge First Certificate in English*™ (FCE®), *Certificate of Proficiency in English*™ (CPE®) papers, and the paper-based TOEFL. However, the FCE and CPE tests have been supplanted by IELTS as the principal means of determining university admission in the United Kingdom. The study therefore has historical interest in terms of examining test delivery, and intrinsic interest in terms of the methodology of the study used, but little practical value in terms of comparing decisions made on the basis of test scores resulting from these instruments.

The authors produced a framework of test method facets as a means of elucidating specific language features that are measured at the item level. Such a process inherently involves both a factual and an abstract component: factual, insofar that instances of specific words (function or content), number of clauses and sentences may be counted and objectively stated; and abstract in that *individual skills* to be measured within each item are unobservable, and therefore their significance is inferential. Therefore, in analysing skills associated with test items, the researchers attempted to make the process more reliable by devising a quantifiable rating system for experienced language practitioners. Their role was to rate the items to determine whether particular skills are represented by particular items (Bachman et al, 1995: 100-105). The authors devised a five-point rating scale to be used to analyse test items for twelve highlighted components of communicative language ability (CLA). The rating scale incorporated considerations of whether or not the skill in question was involved, and if so, to what extent:

Not required	Somewhat involved	Critical Basic	Critical Intermediate	Critical Advanced
0	1	2	3	4

Table 2.1. Five-point CLA rating instrument (Bachman et al, 1995: 102)

Whereas the CLA scale accounted for the various skills and attributes 'encoded' within each item, *test method facets* (TMF) dealt more explicitly with features of the *input.* Thus, the TMF rating scale is concerned with linguistic features that are determined to portray specific skills. For this reason, features subject to TMF analysis include categories such as 'grammar' and 'context'. These were recorded on a rating scale from 0-2 (abstract-concrete):

Abstract	0	1	2	Concrete
Negative	0	1	2	Positive
Counterfactual	0	1	2	Factual

Table 2.2. Three-point TMF rating instrument (Bachman et al, 1995: 201)

The authors' extension of enquiry beyond an analysis of test scores is necessary for two principal reasons. Firstly, results obtained by participants will depend on a broad range of interacting factors that in turn affect test performance, such as the organisation, content, format and presentation of prompt material. The extent to which these factors correspond are important considerations for researchers seeking to compare tests. Secondly, establishing the similarity of the language skills across tests is an important step in making a *validity claim* that the tests that are the object of inquiry measure the same construct. Therefore, to establish claims of test comparability, a focus on construct definition is as important as the technical aspects of score comparability. This importance is further highlighted by examining the literature on score comparability produced more recently by test developers.

ETS (2010), for example, examined the score reports of 1,153 participants who had taken both the TOEFL iBT and IELTS. The lack of item-level data resulted in an equipercentile method being adopted. The focus of the research was on individual sections (reading, speaking, writing and listening) and overall combined scores. Scores were also correlated to determine the strength of relationship between the test components. Results were reported in a concordance table; a range of TOEFL scores were reported for each IELTS band scale for individual sections and the overall score. This was a purely technical exercise that did not consider how each of the four skills

were measured and what impact the different item typed had on the data they gathered. This research is also dependent on institutional support to collect this data and even with this advantage, the datasets were incomplete. The research was designed based on the immediate commercial need to compare IELTS and TOEFL, which accounts for the lack of sufficient item level data available for this research.

Pearson (2009) produced comparability research which was closer to that of Bachman et al (1995). Expert ratings were used to estimate concordance with the CEFR, and a dataset of 26,000 ratings across more than 100 items resulted in the reported concordance. ETS (2007) also attempted to map the TOEFL iBT to the CEFR, again using expert raters. Additionally, as part of the field testing for the Pearson PTE Academic, test takers were encouraged to report other measures of their English proficiency, either self-reported data or official score reports (Pearson, 2009). 2,436 participants reported IELTS scores and a further 144 reported TOEFL iBT scores. Reported scores were used to estimated relationships between the tests, and correlations with the TOEFL iBT and IELTS were reported as .75 and .73 respectively (Pearson, 2009). These reported global correlations were significant at the .01 level, but no data for each of the skill sections were reported. This research was predicted on releasing into the public domain only enough information for test takers to compare scores against each other and the CEFR, but not sufficient data to critically engage with the research, highlighting a weakness of an institutionally-led approach to test comparability.

Engagement with comparability literature has demonstrated that this research is either designed post-hoc to address commercial needs (ETS, 2010), is opaque (Pearson, 2009) and only one comprehensive study exists which addresses construct definitions, albeit for tests which have subsequently been replaced for testing English for academic purposes (Bachman et al, 1995). The question therefore arose of how to obtain information regarding the constructs for the two tests which could be compared. Based upon the engagement with the literature to this point, it was readily apparent that attempting to define the constructs for all four skills would be beyond the scope of this thesis. As outlined, this thesis proposes to concentrate on the skill of *reading*. The 'construct' specifically refers to the operational definition of reading

being tested in IELTS and TOEFL, which is often contained in a working document called a 'test specification'. The next section explores test specifications as understood in the literature, and the extent to which they contain information pertinent to a construct definition.

2.3. Test specifications as realisations of test construct definitions

Fulcher and Davidson (2007: 377) define test specifications as a "generative blueprint"; a document that informs test design and content, and may be used iteratively in the creation of carefully-designed test items, and are designed to continually foster dialogue and debate. The authors also note that there is no agreed-upon model for the production of specifications at any level of granularity (2007: 312). Most authors appear to favour specifications which are linked to the specific audience who will read and engage with them.

Alderson et al (1995: 11-17) distinguish between different types of specification depending on their use; specifications for test validators will necessarily differ from those used by item writers to create new versions of the test. These are typically called item-writer guidelines (IWG). Carr (2011) also identifies different types of specification depending on the audience. He divides test specifications into three sections; test content and purpose, test structure and test tasks. All of these are may be included in a grand 'table of specifications' (Davidson and Lynch, 2002: 66), or they may operate independently for specific stakeholders. Specifications should therefore target the specific context for which they will be used. Specifications may also contain construct information. Buck (2001) outlines two specific methods of defining a construct. A construct may be defined in terms of the tasks that we think a test taker should be capable of performing in the relevant domain, or to state the competencies (knowledge, skills and abilities) that the test taker should possess to function as an active member of the domain.

Specifications which outline test structure and test tasks will resemble item-writer guidelines which outline task parameters so that item types may be replicated and

arranged to create parallel forms of a test or to add items to a battery. Specifications which outline who the test is for, what decisions the results will be used for and how scores relate to claims about aptitude in a construct are of primary utility for the test developers themselves and those interested in establishing a validity argument (Alderson et al, 1995: 9-10). Alderson et al recognise that specification documents may be valuable in articulating (possibly unspecified) theories of language ability or constructs that underlie different item types: "every test is an operationalisation of some beliefs about language, whether the constructor refers to an explicit model or relies upon intuition" (1995: 16-17). More detailed specifications are typically associated with criterion-referenced tests (CRT) in which individuals are determined to be proficient against a list of pre-defined construct-relevant proficiencies (Davidson and Lynch, 2002: 11). Detailed task and item descriptions provide a valuable audit trail to link test construction to these proficiencies.

The most comprehensive set of guidelines for outlining specification content are offered by Davidson and Lynch (2002). The authors propose a model of test specification content based on that of Popham (1978), distinguishing five key components:

1.) General description – a detailed description of what is to be tested, to convey the purpose and motivation of the test.
2.) Prompt Attribute (stimulus) – detailed description of what is to be given to the test taker; what they will see on their screen or test paper. It is this 'stimulus' that triggers the response that is to be measured.
3.) Response Attribute – this describes what the test taker will do; this may be a selected (e.g. multiple-choice selection of an option) or a constructed response (elaborate writing assignment).
4.) Sample Item – this provides a manifest example of the three previous sections and "brings to life" (Davidson and Lynch, 2002: 26) the three previous components.

5.) Specification Supplement – any additional information that is not contained in the previous sections. This may require extra information about the types of text that are to be selected, for example.

This model is primarily descriptive, outlining what the test taker will see and what task they should perform in response to the stimuli presented to them. It is therefore most closely associated with IWG used to create parallel forms of a test. No specific information regarding the construct is cited in Davidson and Lynch's model beyond the general description invocation of 'what is to be tested'.

There is no clear-cut recipe for what a specification should include, with alternative models proposed by different scholars over time. A specification which includes a carefully-worded definition of the construct is not necessarily more evolved than one which does not, as test developers may have decided to publish such information in an alternative document. However, if a specific construct definition includes a particular model of language acquisition, then it would be wise to include this model in the specification, as the understanding of this model and how it is subsequently operationalised would represent a considerable step in the test development process. Stakeholders need to know how test developers' understanding of the construct has informed choices about test construction and how they may make decisions about test use that cohere with this understanding of language proficiency.

Unfortunately for stakeholders, test specifications are an example of propriety information (i.e. they cannot be released into the public domain) (Davidson, 2012). Specifications are rarely released to the general public for stakeholder consumption, and those that are inevitably differ from the internal specification used for the construction of new items. This represents a quandary for stakeholders. They are reliant upon publicly released information regarding the test in order to make decisions about whether to use a test for specific purposes, or if faced with more than one option, which test is most appropriate for their own testing context. When presented with alternative tests which purport to measure the same construct, stakeholders will struggle to determine which test is most appropriate for their own

educational context without access to the construct definition of respective test developers. However, one aspect of the literature that potentially can address this difficulty is the concept of RE.

2.4. Developing a framework of reverse engineering as a solution to the problem of limited information regarding test construct definitions

RE is defined as "the creation of a test specification from representative sample items" (Davidson and Lynch, 2002: 41). This definition implies detailed analysis of representative tasks in an attempt to elucidate common features and provide the standards for each task type. This suggests that RE may be conducted for situations in which a test is created without a specification, so one is inferred, post-hoc, from existing instruments. Additionally, Fulcher and Davidson (2007: 57) define RE as an "analytical process of *test creation* that begins with an actual test question and infers the guiding language that drives it". This latter definition conceives of RE as contributing directly to test development as part of an iterative development cycle. These two definitions illuminate differences in both scale and purpose for which RE may be conducted. That is, RE may be conducted at *item* level or the *test* level, and the rationale for RE may be for the production of new test forms, or for updating a pre-existing specification. The differing agendas inherent in these definitions would produce different research programmes, methodologies and outcomes, yet both could be broadly defined as RE. RE may therefore be useful for the current research agenda of revealing the construct underlying IELTS and TOEFL.

Sections 2.4.1 – 2.4.3 critically engage with the language testing literature on RE to date, finding that the concept has been substantially under-theorised. Section 2.4.2 therefore considers what may be learned from a discipline in which the concept has been more theoretically developed – computer science. The review argues that the understanding of RE in computer science is theoretically useful for developing a framework of RE in language testing, although the framework needs to be further informed from a test development perspective. Section 2.4.3 therefore completes the

framework using elements from *evidence-centred design* (ECD), and argues that these elements can address a range of research agendas identified in section 2.4.1.

2.4.1. The under-developed concept of reverse engineering in the language testing literature

Fulcher and Davidson (2007: 57-58) go some distance towards identifying a typology of different research agendas which may be conducted under the umbrella term of RE. The authors claim to identify five types of RE, each having a distinct research focus although not necessarily conducted exclusively of another type. However, this typology of RE proposed by the authors is better clarified as *two* distinct 'types' of RE rather than five, with the three other types better conceived as broad research agendas of how each of these types may be successfully used to conduct RE research:

Research Agenda	**'Type' of RE**	**Research focus**	
		Object of inquiry	**Aims**
Reliability	Straight	Item(s)	Infer guiding language to (re)produce equivalent items.
Validation	Critical	Item(s)	Critical reflection: are we testing what we want?
Historical	Straight/ Critical	Test versions	To understand how and why the tests changed over time.
Test deconstruction	Straight	Item(s); test versions	Infer guiding language (beyond the item level). Build complete specification to produce equivalent test versions.
	Critical	Item(s); test versions	Examining test design to critically scrutinise: a. Test purpose b. Test developers' interpretation of the domain of interest.
Parallel	Straight/ Critical	Item(s); external standards	Obtaining sample items from multiple sources; inferring guiding language about each sample to determine the extent to which these samples have a similar conception of the domain of interest.

Table 2.3. Typology of reverse engineering (adapted from Fulcher and Davidson, 2007: 57-58).

Fulcher and Davidson's five types of RE are 'straight', 'critical', 'historical', 'test deconstruction' and 'parallel'. In table 2.3, the latter three are represented as research agendas rather than types of RE. Each of these research agendas may be realised using either straight or critical RE, which are clarified here as the two types of RE. *Straight* RE is a procedure of inference described above in relation to Davidson and Lynch's (2002) definition. An item is studied in order to identify its design principles. This process is expanded to other items of the same type. The guiding language is amended as each item is studied until systemic guidelines are able to account for all of the variation within that item type. These descriptive guidelines may then be used to create additional items of that type to place within a test battery. This procedure may be used as part of a reliability argument to show how consistency has been maintained across test forms. *Critical* RE, which goes beyond inference of guiding language, is the process of analysing items to determine whether they measure an aspect (or aspects) of the target domain that we require them to. This is crucial *validity* evidence – that the test measures relevant aspects of the construct. In essence, these are the two types of RE that Fulcher and Davidson refer to.

Historical RE is conducted across multiple iterations of a test, tracing the developmental changes that have occurred. Straight historical RE is conducted to account for design changes. Critical historical RE is conducted to identify how the changing understanding of the conception of the construct drives design changes. Parallel RE is the simultaneous deconstruction of two or more tests. Straight parallel RE is used to infer guiding language to establish grounds for comparison across different tests written to the same specifications to examine how those guidelines have been interpreted by item writers. Critical parallel RE is the process of comparing tests that claim to measure the same construct of interest. This process illuminates how the construct has been realised by different test developers. It is this latter form of RE (critical parallel) that this thesis is concerned with.

There are very few examples of researchers explicitly claiming they are conducting RE. Perea (2011) provides a lone voice. He takes a test-level approach and offers prescriptive guidelines for conducting RE based on the work of Fulcher and Davidson

(2007) and Davidson and Lynch (2002). He adopted a mixed-methods ethnographic case study with evaluative features (2011: 36) by incorporating teachers' reflective feedback on test item format. This procedure was mapped to the five key components of a specification outlined by Davidson and Lynch (2002, see section 2.3).

Perea's approach emphasises a 'critical space' for the production of item guidelines, not only for reproduction of items, but also for reflecting on how the items represent the construct of interest, thus embodying both straight and critical aspects. It also identifies practicing teachers as a key stakeholder group in the production of test specifications in addition to test developers. Perea adopted an ethnographic approach to ameliorate for the lack of methodological and theoretical rigour of RE. Perea's approach demonstrates that different procedures or methodologies may need to be followed depending on the type of RE (straight or critical), or the scale at which it is being undertaken. However, Perea does not critically engage with Fulcher and Davidson's typology to determine which methodologies are most appropriate for different types of RE, as this was not part of his research agenda. Perea's approach, although innovative, demonstrates the paucity of attention that the concept of RE has received in the language testing literature. Specifically, the distinction between straight and critical RE has not been further developed in the literature. A focus on *critical* RE was considered essential for the purposes of this thesis, so this distinction was further explored from the point of view of an alternative discipline – computer science.

2.4.2. Developing the concept of reverse engineering using computer science literature

In order to further develop a framework of RE, it was necessary to explore literature outside of language testing in order to bolster the sparse literature on RE in language testing. This section details the understanding of RE in computer science and how this theoretical literature contributed to the developing framework of RE. RE is a technological concept, with the broad implication of dismantling a machine in order to learn how it functions. It is possible that the concept has not received greater

attention in language testing due to negative connotations of industrial espionage and violation of intellectual property rights, or because it may be practised by rival companies to maintain a competitive position. But it is not the process or concept of RE that is quintessentially good or bad, but the *purpose* for which that process is used which should be used to judge its validity. The position of this thesis is that RE as a concept will have a beneficial outcome for stakeholders.

RE in software development came of age in 1990 (Eilam, 2005: viii) through the publication of an RE taxonomy (Chikofsky and Cross, 1990) and the inauguration of the Working Conference on Reverse Engineering (WCRE). The earliest definition of RE in computer science is offered by Rekoff:

> "the process of developing a set of specifications for a complex hardware system by an orderly examination of the specimens of that system… conducted by someone other than the developer without the benefit of the original drawings… for the purpose of making a clone of the original hardware system" (Rekoff, 1985: 244-252, in Chikofsky and Cross, 1990: 13-14).

Applied to hardware, this definition is meant as an effort to duplicate an existing system (cloning), whereas at the level of software, the aim of RE is to "gain a sufficient design-level understanding to aid maintenance, strengthen enhancement, or support replacement" (surrogacy) (Chikofsky and Cross, 1990: 14), a difference that clearly parallels the distinction between the two definitions of RE offered by Davidson and Lynch (op. cit.) and Fulcher and Davidson (2002). The extent to which RE will be conducted depends upon the original research questions that precipitated the RE. This is evident in the diagram of RE proposed by Chikofsky and Cross (1990):

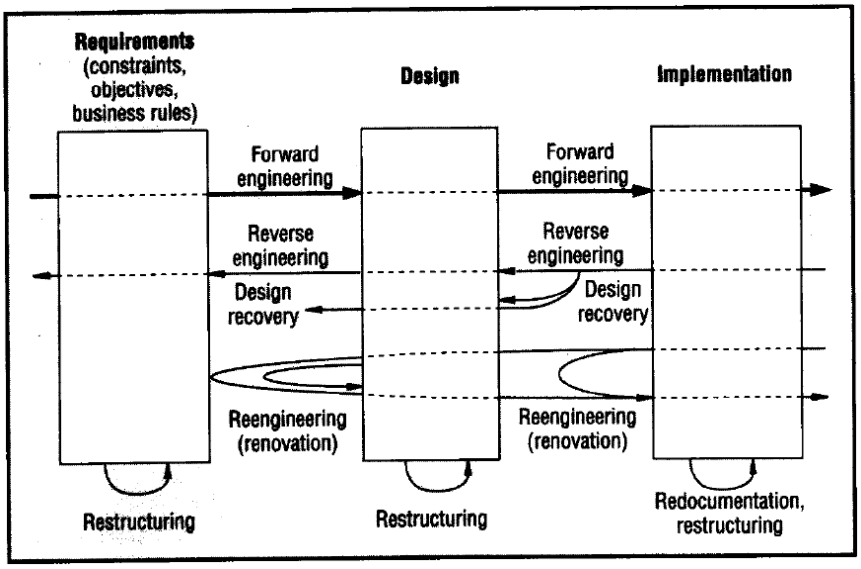


Figure 2.1. Reverse engineering and related processes (Chikofsky and Cross, 1990: 14)

The skeleton from figure 2.1 is extremely useful for the present study. It clearly delineates types of RE for different purposes depending on intended research outcomes, and contains three columns that depict relevant stages of product development which align to procedures of test development. The model also helpfully distinguishes between RE and *retrofitting*, a procedure which has received theoretical attention in the literature (Fulcher and Davidson, 2009).

RE and re-engineering are described as transformative processes occurring "between or within abstraction levels" (Chikofsky and Cross, 1990: 14) in the columns depicted in figure 2.1. For Rekoff, the difference between cloning and surrogacy is not the process itself, but the extent to which one conducts the RE process (Rekoff, 1985: 244). Cloning only requires progressing from the right-hand to the central column of the model. This depicts a process of identification of the design elements of the product, in order to reproduce similar products created to the same specification. Surrogacy, as the process of replacing a product with a new version, progresses from the right-hand column to the left-hand column in order to identify the conditions and constraints in which the product was initially developed. This is analogous to a *mandate* in language testing – the legal considerations and political environment test developers must consider in the production of their tests.

Rekoff cites RE as conducted by researchers not involved in the original development process, but nevertheless significant stakeholders who do not benefit from insider knowledge or the original design templates/blueprints. Conducted by external stakeholders, RE is subsequently assumed to be for the purpose of replication of a competitor's product to allow direct comparison with their own. Thus the distinction between *RE* and *retrofitting* as defined by Fulcher and Davidson (2009: 124) is that the latter is explicitly conducted by original developers in relation to explicitly stated requirements. *Upgrade retrofitting* realigns the test instrument with its originally-stated purpose. *Change retrofitting* refers to modifying an instrument to meet the needs of a different test purpose (Rekoff, 1985: 124).

Although the skeleton offers a useful *outline* for a framework of RE, it was clear that to develop comprehensive framework, it was necessary to demonstrate how the skeleton related to test development. The next section demonstrates how the componential approach to test development embodied in evidence-centred design (ECD) forms the content of the final RE framework.

2.4.3. Developing a comprehensive framework of reverse engineering using evidence-centred design

The approach to conceptualising RE within an explicit framework was inspired by Fulcher and Davidson's (2009) use of the principles of architectural layering to reconceptualise the process of 'retrofitting' a test (e.g. to meet a new purpose). The notion of architectural layers in language testing was formulated by Fulcher (2009), who interprets the documentation of test preparation as distinct layers. Layers are "modular and independent" (ibid.: 126) and progress from more generalised to more specific contexts. The notion of layering is also clearly pertinent to RE, as the discussion of RE in the context of computer science above attests.

This section demonstrates the utility of the modular approach of ECD to developing a RE framework. It begins with section 2.4.3.1., outlining the relevant elements of ECD,

called the *conceptual assessment framework* (CAF). Sections 2.4.3.2. outlines the essential theoretical requirements inherent in the CAF which the RE framework must incorporate, namely how the concept of *validity* is understood in ECD. Section 2.4.3.3 outlines the final RE framework as understood from computer science literature and ECD.

2.4.3.1. Relevant elements of evidence-centred design to the framework of reverse engineering

Mislevy et al (2003) and Mislevy and Riconscente (2005) demonstrate how the concept of 'architectural test layers' functions in theory and practice. This is termed 'evidence-centred design' (ECD). ECD was proposed as a means of formalising the test design process and drawing relations between the individual stages of test design for designers and item writers (Almond, Steinberg and Mislevy, 2002; Mislevy, Almond and Lukas, 2003; Mislevy and Yin, 2012). As ECD attempts to provide a comprehensive framework for test development, relevant elements could therefore be applied to an existing instrument to provide an efficient and principled approach to RE that could encompass a variety of research agendas. The most crucial components of ECD for this thesis are outlined in the conceptual assessment framework (CAF). Fulcher and Davidson (2009) align the CAF with their metaphorical architectural layers of a test, providing a precedent for applying the CAF to retrospective tests analysis in order to consider whether the final program ('as built') conforms to the architectural design ('build as'). This section therefore explicitly outlines the CAF and considers the utility of the CAF as the basis of a framework of RE.

The conceptual assessment framework includes six models. Three of these are integral to a descriptive test specification developed through straight RE by Fulcher and Davidson (2009: 127); the *presentation*, *assembly* and *delivery* models. The remaining three will be examined regarding their utility for the development of a construct-based test specification created through *critical* RE; the *student*, *evidence* and *task* models.

The presentation model provides information regarding the organisation of the test; instructions for how items and relevant material should be organised either on paper or on screen. The assembly model describes how the student, evidence and task models cohere to provide sufficient construct-representative information. For example, if the developers claim that each item targets some aspect of the construct, then the assembly model specifies the number of each item type to ensure that the construct is adequately represented. The assembly model also provides guidelines on the content of accompanying material (e.g. reading passages). In computer-adaptive testing, where the items presented to test takers depend on test taker responses (if test takers consistently answer incorrectly, they are progressively presented with easier items, and vice versa), a test may not have an explicit assembly model. The delivery model describes how all of the constituent models function together to deliver the assessment. This includes information on security, relevant platforms, timings (Mislevy et al, 2003).

The CAF is organised around the three core models (student, evidence, task) which are vital to the assessment argument (Mislevy and Riconscente, 2005: 17) and are the central models of task design (Mislevy and Haertal, 2006: 10). The relationship between the six constituent models is outlined in figure 2.2 below, and a detailed description of the three core models follows.

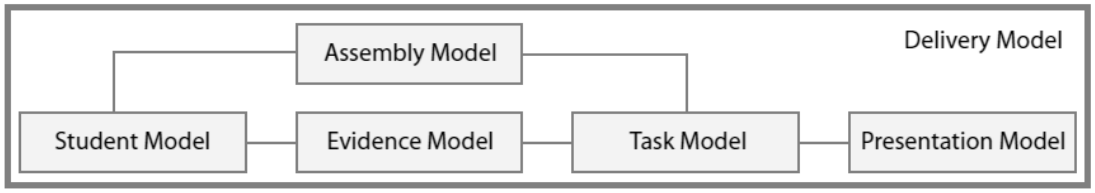


Figure 2.2. Design models of the ECD Conceptual Assessment Framework (Almond et al, 2002: 12).

The *student mode*l (SM) describes the construct (Fulcher and Davidson, op. cit.: 128) in terms of what is being measured (knowledge, skills and abilities). There may be one or more variables in the student model. Measurement of the student model may vary

from a straightforward assignation of a dichotomous integer score in response to an item to a more complex multivariate probability distribution built in response to multiple, partial-credit items (Mislevy et al, 2003: 6; Mislevy and Riconscente, 2005: 17). The student model contains *student model variables*, which operationalize the knowledge, skills and abilities identified in domain modelling. SM variables provide the link between the assessment performance and subsequent claims (Mislevy and Riconscente, 2005: 17) about test takers. The student model should detail the level of granularity of the student model variables, each of which may be operationalized through different items and item types. Thus each response provides further information to a final student model which gives a 'picture' of the proficiency of an individual test taker.

The *evidence model* (EM) details the evidence required to make inferences from *observable variables* (work products) to the construct of interest (Fulcher and Davidson, 2009: 128). Each instance of student behaviour provides evidence of one or more of the student model variables. As more information is gleaned, a fuller picture of the student is created. Evidence rules detail how the work product is interpreted and translated into a score (Mislevy et al, 2003: 12). In items for which the work product is the selection of a single option in a multiple-choice question, the psychological interpretation of the item on the part of the task designer informs the *evaluation* component of the evidence model (Mislevy and Riconscente, 2005: 18). A measurement component details how the student model variables are accumulated and contribute to the scoring procedure (ibid.: 12).

The *task model* (TM) is a schema (Almond et. al., 2002: 491) which describes the goal-directed activities (Mislevy and Haertal, 2006: 5) or environment (Mislevy et al, 2000: 8) in which test takers produce the work products for evaluation. They describe the material that is presented to the test takers and what the expected work product will look like. *Task model variables* detail how task features correspond to work products. Different task models will relate to different task types. These task models are used to control the evidence that emerges in the form of work products (Mislevy et. al., 2003.: 14). Separating the task model from the evidence model allows developers to define

scoring independently of task design, potentially allowing tasks to be used for different assessment purposes or in different assessments (Mislevy et al, 2000: 6).

The major advantage of adopting the CAF framework is the componential approach to specification development that has been outlined in this section. Different test developers or stakeholders conducting RE will require different specifications depending on their purpose or research questions. The CAF is sufficiently flexible to allow for various approaches to RE at different levels of granularity. However, before specific research questions can be formulated for the current study, the implications of adopting an ECD approach need to be further explored.

Any framework for RE built on the CAF must satisfy three conditions inherent to ECD. It must have a "level of generality that supports a broad range of assessment types" (Mislevy et al, 2003: 3). In other words, the framework must be flexible enough to be applied to a variety of research purposes, situations and contexts. Secondly, the framework must consist of individual models that are as rigorously defined, as each model contains an intrinsic logic and structure (Mislevy and Haertal, 2006). To avoid 'leakage' within the CAF, each model must be able to be defined independently of other models so they may be readily operationalised according to the research goals of the reverse engineers. Thirdly, the framework must have a very clear evidential basis and satisfy Mislevy's process of "evidentiary reasoning" (Mislevy et al, 2003: 3) in order to directly link claims about tests to collected data, for example, by using Toulmin's structure of argumentation (2003). As this thesis adopts the three core models of the CAF as a framework for RE, this framework will also be dependent on this process of evidentiary reasoning, which forms the conception of *validity* in the current study. The next section will therefore briefly outline the concept of validity as understood in the language testing literature, before looking more closely at the nature of validity inherent in the CAF and the implications of this for a critical parallel RE study.

2.4.3.2. The concept of validity in the language testing literature

Test validity is defined as “an integrated evaluative judgement of the degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness of inferences and actions based on test scores or other modes of assessment” (Messick, 1989: 13) This definition contains four key components (Chapelle, 2012: 22), all of which are relevant to the current study.

First, validity is a property of the inferences made on the basis of information (scores) collected via a test, rather than a property of a test itself. Kane (1992: 527) states that validity as a concept refers to “interpretations of test scores and the appropriateness of those interpretations”. This is known as an interpretive approach to validity and underpins the approach to validity inherent in the CAF. This is explicitly outlined in the next section.

Second, validity is best thought of as a single unitary concept, with different stages of test development contributing to a whole, progressive argument. Davies (2012: 39) distinguishes between a process of *validation* as distinct from the abstract notion of *validity*. Kane, in the interpretive approach to validity, is concerned with a logical validation argument that directly links decisions made about individuals with the test developers’ conception of the construct through a linked argument and evidential audit trail.

Third, validation includes wider social consequences associated with high-stakes testing. Messick’s definition was instrumental in advancing a notion of ‘consequential validity’; that the intended and actual outcomes of a test be part of a validity argument. Bachman and Palmer (2010) incorporated the concept of consequential validity in an *assessment use argument* (AUA) framework in which the intended use of an assessment must be outlined and justified as an intrinsic part of test development (Bachman and Palmer, 2010: 94). To date, the evidential basis of *test use* has formed the mandate for test comparability research. That is, if two tests are used for the same purpose (such as access to higher education in an English-speaking country), a mandate is created to compare those tests statistically to identify whether decisions about individuals are consistent across instruments. However, the position of this

thesis is that claims linking a score on one test to a score on another test depend not only on aligning scoring rubrics, but also require questioning whether the test developers have a shared conception of the *construct*.

Fourth, validation is an ongoing, recursive process. It does not end when a test goes 'live'. It must be noted that studies linking scores across instruments cannot form part of the validation argument for either test, as score comparability is a technical exercise unrelated to the psychological aspects of the competence that either test is designed to reveal. Test comparability can only succeed if validation arguments withstand scrutiny according to predefined standards of acceptability for a validation argument. Therefore, test comparability research must be concerned with *construct definition* that underpins the assessments. Construct-relevant data collected as part of a comparability exercise therefore *can* form part of a validation argument. Defining the construct under investigation is a key part of test development that underpins item design and creation, and is also crucial to critical parallel RE. *Construct validity* is therefore central to this thesis. Fulcher (1999: 226) defines construct validity as "the extent to which the content of the items included in the test can be accounted for in terms of the trait believed to be measured". This definition directly links content and construct; the former being the embodiment of the test developers' conception of the latter.

2.4.3.3. The argument-based approach to validity inherent to ECD

The approach to validity in the CAF as outlined in the previous section is embodied in Kane's (2006) adoption of Toulmin's (2003) interpretive structure for argument, specifically termed a 'validity argument'. Interpretive arguments are highly versatile. They can span an entire assessment framework (Fulcher and Davidson, 2009: 134) which determines how individual architectural layers are linked, or can operate at the level of items or tasks (Fulcher and Davidson, 2007: 162-175) to determine how a claim regarding an item design can be linked to a correct score being awarded for that item. The interpretive argument structure can operate successfully as a validity argument for

test construction (Kane, 2006; Chapelle et al 2010) and is therefore suitable for a RE framework.

Kane articulated a 'network of inferences' (2006) from test score to claim regarding a test taker's language ability. The epistemological framework for this process of argumentation is outlined below:

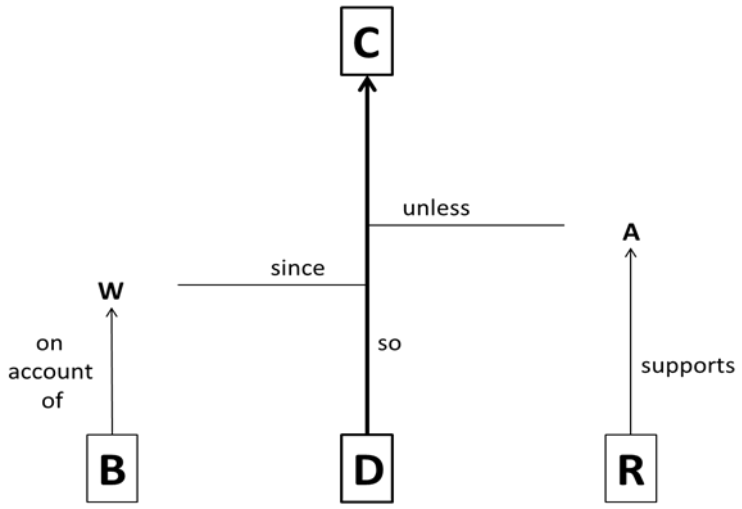


Figure. 2.3. Toulmin's structure for arguments. Reproduced in Kane (2012: 209)

The general structure of a Toulmin argument is composed of a *claim* (C), or statement that we wish to make. This claim is predicated on *data* (D); 'facts' which are used in support of a claim. A *warrant* (W) is any statement that supports the claim. Warrants require evidence in the form of *backing* (B). Claims may be challenged by *alternative explanations* (A), which is supported by *rebuttals* (R); statements identifying conditions in which the warrant is not relevant, or is contradicted. Chapelle et al (2010: 5) compares this approach to a legal argument, in that lawyers must convince judges or juries of the innocence or culpability of a client or defendant on the strength of their assessment use argument (Bachman and Palmer, 2010: 94).

Chapelle (2010: 5-6) outlined a series of propositions of the types of claims that could be made in relation to score interpretations – these propositions (claims) require a

diverse range of evidence to provide backing. These propositions consist of positive statements related to acquired language skills, the context in which the test is taken, reliability, construct-irrelevant variance, scoring procedures, measurement and scoring rubrics, test bias, a model of language 'ability' and washback implications of the assessment. The diversity of these statements demonstrates the array of topics that necessarily form parts of a whole validation argument, and therefore the flexibility of the interpretive approach to validity.

However, Kane's argument-based approach to validation has received criticism. Davies (2012) argues that Kane conflates the practical approach of 'validation' with the abstract 'validity'. As the argument-based approach does not produce a workable formula by which future tests or test developers may use this approach. As a result, it is unclear what type of evidence (*data*) is required that will adequately stand up to scrutiny in the form of a *claim* about test taker performance. Oller (2012) makes a similar criticism, that Kane's invocation of a legalistic argument to state that such logical reasoning requires *groundedness*, and that the epistemological underpinnings of an argument that centralises interpretivist thinking requires direct linkage to concrete *true narrative representations* (TNR) from which we may generalise up to the "limit of similarities" (ibid.: 32). This criticism offers both a strength and a weakness for the current study. A strength, in that interpretive arguments are not over-prescribed, and offer independent researcher scope in which to develop their own arguments based on their own research agenda. A weakness, in that this study therefore requires considerable methodological emphasis to determine what evidence (data) should be collected as part of the current research agenda.

The general framework of RE in this study is predicated on the CAF, and is therefore grounded in interpretivist epistemology. This fits well. Claims made about the nature of the construct understood by the test developers can only be made on the basis of the information collated about the tests in the public domain. RE research based on a limited sample of test materials may only present a limited picture of the domain. However, Oller's requirement of the 'limit of similarities' provides the basis for generalizable claims associated with RE outcomes. If the RE instruments are written

from the same specification as live materials produced by the developers, then the research instruments exist within the limit of similarities proposed by Oller. This is similar to a form of *analytic* generalisation, which is discussed further in the methodology.

2.4.3.4. An evidence-centred framework of reverse engineering

Based on the discussions of RE in computer science and language testing, evidence-centred design (specifically the conceptual assessment framework) and validity in the previous sections, this section now outlines the emergent framework of RE and provides commentary of how the relevant components emerged from the literature discussed. The framework is represented in figure 2.4 below:

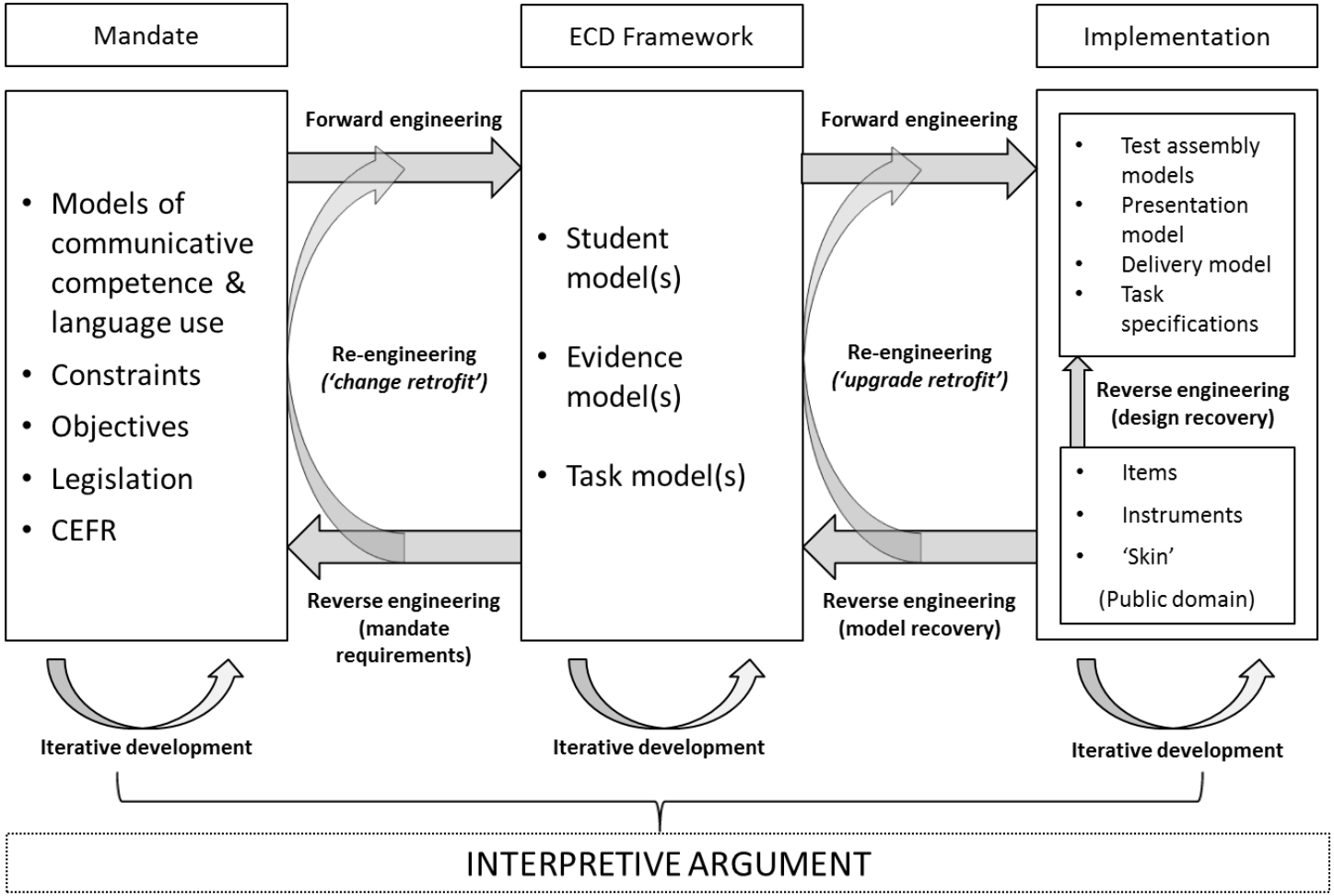


Figure 2.4. An evidence-centred framework of reverse engineering.

The framework is inspired by the framework of RE in computer science detailed in figure 2.1. Figure 2.1 clarifies the difference between retrofitting and RE, a distinction which has not been elaborated in the language testing literature to date. It also demonstrates that RE can occur at any stage pf the development cycle, and can uncover criterial design features or the constraints in which the relevant product was manufactured. Figure 2.4 demonstrates that many of the principles of RE in computer science are readily applicable in language testing. This section will outline the contents of the proposed framework.

It contains three columns. From left to right they represent traditional broad stages of 'forward engineering'. An institution, government agency or other organisation identifies a need to make decisions about individuals related to some aspect of language proficiency and decides that an objective assessment acting in accordance with existing legislation and best linguistic theory undertakes assessment design, construction, field testing and then launches the test into the public domain (mandate, ECD framework and implementation). Each of these stages includes iterative development by the developers to improve the coverage of the target domain or improve statistical qualities or strengthen validation arguments. The question underlying the evidence-centred approach to RE is**:** how can we make RE more objective? This framework is offered as a means of creating reader/user responsive research outcomes. In conducting RE to make a new specification or to uncover domain-relevant design principles, the engineers will need to consider who the specification is for and why. Different stakeholders will require different information. The goal is to forward a standardised approach to RE to provide strong epistemological grounds for comparing outcomes for different research instruments subjected to RE.

A general framework of RE encompasses both straight RE (the right-hand column) and critical RE (the central column). Within the central column, the student model encompasses the proficiencies which the developers have encoded in their items, and so represents the construct as it has been uncovered in the research project via the evidence and task models. The distinction between straight and critical RE is addressed directly in the framework, the constituents of which are now outlined.

2.4.3.4.1. Mandate

The left-hand column represents the mandate. This is the constellation of forces that shape the design of the test. Critical test deconstruction as expressed in RE attempts to establish the mandate that has shaped each test. This is an attempt to move up/outside the test to elaborate the forces that shaped the test. The form that assessment takes in any society tells us something about how that society values education and the role they believe tests must play in ordering that society. El Atia (2004, 2008) conducted an historical examination of the French baccalaureate and its place within French society and how it was shaped by French protests in the Napoleonic era. She used ideas from critical theory to examine the test in its wider context. This critical approach to testing is further evidenced in Shohamy (2001) who examined the power of tests in their social context.

The mandate therefore refers to the *conditions* under which the test was conceived and constructed. Discourse between and among stakeholders and test developers necessarily occurs under these conditions to determine the *objectives* of the new test. How these conditions developed or are controlled in the discourse which results in the production of a test will affect the consequential validity of the test; that is, the effect of the test on test takers and wider society. Predicting or identifying of the effects of the test on different test taker populations may place *constraints* on the test is able to measure. Vulnerable groups of test takers may also be protected by national *legislation* to ensure that they are not unintentionally penalised by some aspect of the test taking process which is not construct-relevant.

The Common European Framework of Reference (CEFR) is also part of a mandate. In a situation when a new test is required for university admission purposes, the university hierarchy may insist that the new test must align with the CEFR. An applied linguist may look at the CEFR and imagine archetype items that relate to the descriptors and the language used to describe these items may also occur to the linguist. This is called ‘spawning’ (Davidson, 2015, personal communication). This is the origin of a new

specification. This does not imply that the generation of a new test and specification is a non-creative process. 'Archetypes' and 'spawning' still leave room in the linguist's head for a creative process for new item types and original examples of these items. The specification is in fact the vehicle for creative consensus building and a record for the evolving thought of the test developers. Specifications evolve organically, not only when new written version is introduced, but also when test developers and/or teachers use a test and reflect critically on which tasks performed as expected, which did not, and how these tasks may be improved in the next iteration. This represents a further example of how specifications may be 'spawned'.

The left-hand column also includes a clear definition of competence and/or language use which will undergird test and task design. The use of ECD in test and task design does not preclude the use of linguistic models to inform test creation practice. For example, Bachman's (1990) and Bachman and Palmer's (1996) model of communicative competence is derived from that of Canale and Swain (1980) and Canale (1983), which "subsumes the work of both Chomsky and [Dell] Hymes" (Skehan, 1998: 158). Hymes' contribution was to incorporate the *appropriateness* of utterances. This sociolinguistic aspect of competence then provided a strong foundation for subsequent models in Applied Linguistics and formed part of the mandate for the TOEFL 2000 project, culminating in the TOEFL iBT.

2.4.3.4.2. Design Recovery

The RE process begins in the lower right hand corner of the framework. With the gathering of information related to the test in the public domain, including authentic examples of instruments produced by the test developers that are claimed to be authentic examples written from the specification. In-depth analysis of all observable variables (test format, item types and input material) of the test instruments includes the length of test, characteristics of input material; instructions; presentation and structure of questions and input material and the expected work product. If the test is computer-based, this may also include features of the interface. A taxonomy of features to consider in interface design advanced by Fulcher (2003: 385) includes:

“navigation, text, page layout, terminology, help facilities, icons/graphics, the use of colour, and toolbars.” The extent to which the electronic format affects test-taker performance is a vast area of study in itself.

The evidentiary basis comprises the task paradigm and task, presentation and assembly models. In the latter three stages, we are identifying the salient features of individual test items (e.g. dichotomous or open) and analysing the language used and how they are presented to test takers (pen and paper, electronic, font size, typeface, page layout). As the TOEFL iBT is delivered in an electronic format, consideration of interface design is of importance to this stage of the RE process. This was accounted for in the present study by utilising the forms of test items presented to test takers in paper-based format in preparation materials.

From a reading perspective, RE at this stage considers basic features of the prompt material, which questions were written in relation to them, the explicit wording of the questions, and the expected procedure for successful completion of each item (and the expected ‘work product’ or response of the test taker). For receptive skills such as reading this process is comparatively easier, as the interplay between the input material and the questions inevitably reveals the procedure for procuring the correct answer. However, for productive skills (speaking and writing) there is less interplay for those tests that do not test integrated skills. Reverse engineers may continue this process until each new example of each item type they find is accounted for by the design principles they have inferred. This finalised specification will contain descriptions of the assembly, presentation and delivery models. This is straight RE aimed at design recovery.

2.4.3.4.3. Model Recovery

Critical RE (which may or may not include design recovery as part of the research mandate) uses the publicly available test material to uncover the elements of the ECD framework in the central column. In relation to Mislevy’s ECD, these components include the student, evidence and task models.

2.4.3.4.3.1. The Student Model (SM)

The aim of the *student model* (SM) in RE is to identify *what* is being measured. In straight RE, engagement with the instrument is sufficient to uncover the design principles associated with particular items. In critical RE, researchers need to engage with both the instrument and the test taking population to identify which mental processes are encoded into item design. The complete picture of processes that the test taker reveals is the student model. This process entails identification of which processes are construct-relevant and which are not construct-relevant. Test comparability requires access to the full repertoire of student model variables to make justifiable comparative claims between instruments.

In critical RE, the student model is expressed in terms of the cognitive processes encoded in the individual items that are the product of deeper engagement with individual items types. Descriptive design principles inform the nature of the data collection and identify the salient task features with which the test takers will engage (these are the task model variables). These then form the locus of attention for the test takers and researchers. Individual moments of engagement reveal instances of cognitive processing that the researcher can claim are associated with that item type, if they contribute to the test taker responding correctly to that item. Once complete, the 'cognitive map' accounts for the levels of cognitive processing required to complete specific item types. This is then presented in the form of a matrix, outlining the requirements of the test in cognitive terms. Once elaborated, this becomes the basis of decisions about that test, such as whether is suitable for specific purposes, and whether or not the test is comparable to other tests that claim to sample the same domain of interest.

2.4.3.4.3.2. The Task Model (TM)

In straight RE, elaboration of the *task model* initially involves in-depth analysis of all facets of the test tasks: the number of items; the item types present; why specific texts were chosen, which questions were written in relation to them, the explicit wording of the questions, and the expected procedure for successful completion of each item (and the expected 'work product' or response of the test taker). For receptive skills such as reading this process includes analysis of the relationship between the input material and the questions, revealing the relevant parts of the text for locating the correct answer, and thus, how the skills identified in the domain model have been conceptualised by the item writers. For electronically delivered tests such as TOEFL, descriptive task models also include taxonomy of features to consider in interface design including: "navigation, text, page layout, terminology, help facilities, icons/graphics, the use of colour, and toolbars." (Fulcher 2003: 385).

In critical RE, task models consider how the task model variables link to the expected work product and therefore whether specific tasks are associated with specific claims made about successful test takers. Essentially, this is a critical form of RE conducted at the micro-level within the test to determine what cognitive processes from the student model can be associated with particular tasks. RE at these stages might initially consider why specific texts were chosen, which questions were written in relation to them, the explicit wording of the questions, and how these impact successful completion procedures for each item. Test taker involvement in the project provides substance to inferences made about the processes believed to be involved in item completion. For receptive skills (reading and listening) this process is comparatively easier, as the interplay between the input material and the questions reveals something about the procedure for procuring the correct answer, and thus, how the skills identified in the domain model have been conceptualised by the test writers. These can be confirmed/falsified by observing test takers completing individual items. However, for productive skills (speaking and writing) there is less interplay for those tests that do not test integrated skills, such as IELTS. The production of a specification for a test therefore requires parallel analysis of the questions, work products and the corresponding marks awarded. Analysis of work products at varying levels of competence will allow researchers to determine how the skills from the target domain

are conceptualised for the productive skills at a similar level of granularity to the receptive skills.

2.4.3.4.3.3. The Evidence Model (EM)

In the context of the current research, the *evidence model* (EM) considers what data is needed in order to make claims about individual tests and how this data may be analysed to identify construct-relevant attributes. Researchers undertaking RE therefore need to consider how the representation of the domain model is encoded in specific items, how RE can progress from a publicly-available test instrument to specifically defined competencies that reveal something about the test developers' conception of the target domain. Carefully-defined cognitive processing can then be the "operational definition" of competence (Mislevy, 1992: 23) for both test developers and stakeholders.

The evidence model considers what information is needed to make a judgement about individual test takers and how this evidence relates to the cognitive processes required in the domain. The specification matrix acts as a 'cognitive rubric' so that individual items types may be sampled and fed into an instrument to produce a test that samples the domain of interest. Researchers undertaking RE therefore need to consider how the cognitive processes from the domain model are conceptualised in the items, how broadly defined domain competencies can relate to specifically defined competencies that are revealed by RE. For example, a B2-level reader is described in the CEFR in the following way: "I can read articles and reports concerned with contemporary problems in which the writers adopt particular stances or viewpoints. I can understand contemporary literary prose" (Council of Europe, 2001: 6). A researcher should therefore ask the question 'what utility does this description have for the target domain? What do 'contemporary problems' refer to or 'contemporary literary prose' relate to, in terms of the competencies I have revealed from the RE? It is only different groups of stakeholders within the target domain that can adequately answer these questions. Stakeholders can then make a decision about whether a test which samples

the construct of interest or matches the broad descriptors of the CEFR in such a way as to legitimise the use of a test for their own purposes.

2.4.3.4.3.4. The final outcome of RE: The test specification blueprint

The test specification blueprint should provide a cogent summary of test response analysis, and provide clear evidence for how the cognitive identified as part of the construct have been operationalised. The blueprint may vary in its level of granularity depending on the needs of the researchers or other stakeholders (e.g. straight or critical RE). The specification produced from critical RE should provide a valuable piece of evidence between the observed variables that the test requires for successful completion and the construct. This is why the framework of RE is underpinned by an interpretive argument. A test specification at a fine level of granularity should show how these processes have been operationalised at the item level, to the extent that a researcher may create items that have congruence with existing items (Fulcher and Davidson, 2007: 377) at the *cognitive level*.

The literature review has now outlined the framework of RE that will form the basis of this study, based on literature from language testing, computer science and evidence-centred test development. As this study proposes to adopt a *cognitive* approach to comparing the IELTS and TOEFL reading tests, the remainder of the literature review focuses specifically on current understanding of cognitive approaches to studying reading in test-taking environments, and how evidence of cognitive processing in reading tests has been captured methodologically.

2.5. Exploring and understanding second language reading

This second part of the literature review considers how the testing of L2 reading is understood in order for the study to successfully identify constituent parts of the construct embedded in IELTS and TOEFL. Specifically, this section examines recent literature which presents a cognitive perspective of reading. The section also explores the literature regarding how reading is characterised in IELTS and TOEFL. This section also considers *how* cognitive understanding of second language reading is studied, identifying and examining the strengths and weaknesses of verbal protocol analysis. The literature review concludes by integrating the understanding of a cognitive view of reading with the newly-developed RE framework, resulting in the development of four critical research questions, which will guide the data collection and analysis.

2.5.1. Cognitive approaches to studying second language reading

Reading is measured indirectly (Field, 2008: 270) as we infer that correct responses relate directly to unobservable cognitive processing. The test taker is presented with input material, which must be decoded with reference to the participants' lexical bank. Individual phonemes are identified and grouped to form words. Words are then parsed, so that the test taker utilises their grammatical and structural knowledge of the language to organise the input material into coherent semantic units. Organised segments are then combined to identify the ideas and arguments contained within sentences. Arguments (major and minor points) are then structurally organised in the reader's mind to produce an organised mental representation of the whole text.

Reading in a test taking environment is goal-oriented. That is, the purpose of reading is to extract very specific information for the purpose of addressing a particular task or set of tasks. This action will be time-constrained unlike the target domain. As the reader progresses through the text, the mental representation of what has gone before is stored in the reader's working memory. Reading nonetheless entails recursive strategies to allow test takers to reconsider a response by returning to part of the text when their comprehension was not sufficient to confidently respond to an

item (Field, 2008: 27-28). Thus, researchers must invoke a cognitive argument in order to justify test tasks being used to make high-stakes decisions about test takers which are not representative of the types of tasks they will undertake in the domain.

Khalifa and Weir (2009) proposed a componential model based on a cognitive approach to reading. This model offers four distinct advantages for comparing cognitive processes in reading in IELTS and TOEFL. First, the componential approach makes the model amenable to transformation into a research instrument to be used for coding purposes. Statements regarding how different cognitive processes relate to different item types may therefore be formulated. This gives rise to an additional issue: the use of each component individually for coding purposes requires the consideration of the strength of evidence presented for each component. Second, the model explicitly incorporates cognitive processing occurring at local and global levels. Khalifa and Weir's (2009) model is designed to be hierarchical. Each level is conceived as subsuming the lower levels. Each level represents a 'higher' level of processing as the unit of analysis increases, thus increasing the cognitive load on the reader. Thus, each level is conceived as being more difficult than the previous levels. Third, the model has successfully been used in a variety of contexts. Weir et al (2000) developed the Advanced Reading *Test* in *English* for Academic Purposes (*AERT*) in China on the basis of this more expansive model. Most usefully for the present study, Weir et al (2009a) used the model to investigate the cognitive processes intrinsic to the academic reading component of the IELTS, using retrospective questionnaires and verbal reports. Fourth, the model derives from in-depth analysis of psychological L1 literature, meaning that it has a strong empirical base. Nonetheless, the model will not be adopted uncritically. This section will explore the different components of the model to further explore its suitability for the current study. The model is outlined in figure 2.5 below.

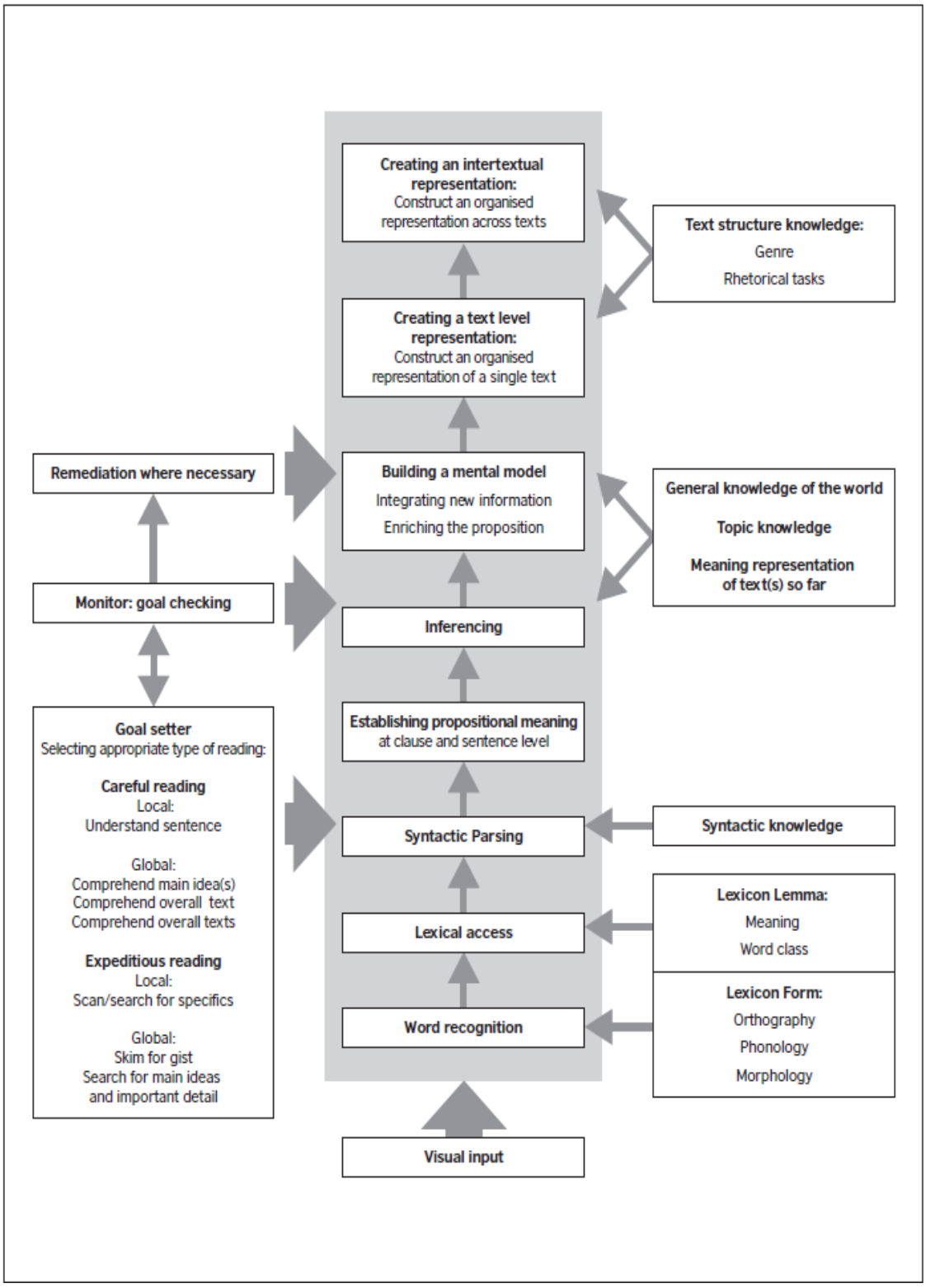


Fig. 2.5. A Cognitive Model of Reading (Khalifa and Weir, 2009)

The model contains three columns. The left-hand side represents *metacognitive activity* – conscious decisions made by the test taker regarding how they approach a text, remedial activity to account for deficiencies in processing and identifying a purpose for reading. The central column represents the cognitive *processing core* – the cognitive processes activated in the action of reading. The right-hand column references monitoring activity. Monitoring refers to the linguistic and world knowledge that the reader brings to bear on the text. The model therefore references both bottom-up and top-down processes. The metacognitive and monitoring columns impact directly on the efficacy of the cognitive reading processes, which begin at the orthographic, phonological and morphological levels in response to text-based input material. The three cores are now discussed in further detail with reference to wider language testing literature in these areas.

2.5.1.1. Metacognition and monitoring

Metacognition broadly implies thinking about thinking. That is, deliberate and goal-oriented deployment of a strategic action which the language learner is aware of (and can report using) and considers pertinent to a specific task (Phakiti, 2003: 29). Nelson and Narens (1990) proposed a model of *metacognition* (knowledge about and regulation of one's thinking) which operates at two identifiable levels: the cognitive (object) level and the metacognitive level. The *cognitive* operations refer to the mental processes that occur in turning ideas into language, or conversely, language into ideas based on the linguistic repository available to the interlocutors (see section 2.6.2). The *metacognitive* operations refer to the higher order control over this procedure; the executive functions that manage the retrieval of linguistic referents for organisation into coherent ideas and grammatically recognisable units.

Bachman (1990) and Bachman and Palmer (1996) echo this distinction by separating strategic competence from language ability. Strategic competence was originally characterised by Canale and Swain (1980) as a compensatory measure to make up for some deficit in linguistic competence. However, Bachman and Palmer's approach is far

broader and incorporates test taking strategies as part of identifying requisite linguistic resources; goal-setting; planning and execution. Strategic competence is therefore a general ability determining the management of available resources (metacognitive decision-making and monitoring strategies).

Bachman and Palmer (2010) argue that overall strategic competence is a manifestation of how metacognitive strategies regulate the cognitive strategies during task completion. Phakiti (2008) further divides metacognitive strategic competence into *strategic knowledge* and *strategic regulation*. This split evokes Chomsky's 'performance-competence' (1965) distinction. Strategic knowledge refers to the set of strategic actions available at one's disposal (knowledge of goal setting, planning and monitoring). Strategic regulation is how one actually employs these faculties in any given setting. 'Strategic competence' therefore refers to a set of metacognitive strategies which regulate the cognitive engagement with a task which operates to produce language either in a productive task or in the mind of the test taker.

Two broad types of metacognitive strategies are identified in Khalifa and Weir's model – planning and monitoring, the former being the previewing of tasks in order to establish the requirements of the task. This refers to the regulation of the test taker's own thought processes in terms of determining which steps need to be taken in order to successfully complete a task. The latter refers to the purposive evaluation of their performance in relation to task requirements. Goal setting in the model refers to conscious decisions by the test taker to search for information either in particular parts of the text (local), or that the task requires an overall understanding of the text (global), which will subsequently influence their observable interaction with the text. Local and global understanding are further modified depending on whether the test taker believe that the necessary information may be gleaned by reading carefully or expeditiously. That is, whether they must carefully parse each lexical unit, or skim the text for specific information.

This four-way model of decision-making was initially proposed by Urquhart and Weir (1998) and can be interpreted as strategic competence. Weir et al (2009a) investigated

strategic processes intrinsic to the academic reading component of the IELTS, using retrospective questionnaires, verbal reports and expert judgement. The authors identified a range of careful and expeditious reading strategies based on the Khalifa and Weir's componential model of reading. Quantitative findings that emerged from the study are presented in table 2.4 below:

	Read expeditiously by			**Read carefully for meaning which is**				**Reading**		
Test taker analyst	skimming	Search reading	scanning	Explicit within sentences	Implicit within sentences	Explicit across sentences	Implicit between sentences	To construct a text model	For a situation model of text and own prior knowledge	Totals per reader analyst
A	0	45	50	277	27	115	45	3	0	**562**
B	70	6	93	318	12	57	25	4	0	**585**
Cognitive skills total	**70**	**51**	**143**	**595**	**39**	**172**	**70**	**7**	**0**	**1154**
Sub-totals: expeditious vs careful reading	**264**			**883**						**1154**

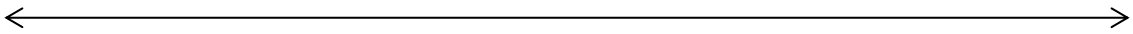
Table 2.4. Summary of responses to two test taker analysts to reading test tasks of 14 authentic IELTS reading tests (Weir et al, 2009a).

The aim of the quantitative element of the study was to differentiate between careful and expeditious reading in the IELTS test. The two analysts reported more than three times as many instances of reading carefully as reading expeditiously. 883 of 1154 reported strategies (76.5 per cent) were instances of participants reading carefully rather than expeditiously, of which 595 (51.6 per cent) were *reading explicitly within sentences*, with only 242 instances of establishing propositional meaning across sentences (21 per cent) (172 explicit across sentences and 70 implicit across sentences). Participants also reported no engagement with the text as a whole for any of the IELTS questions, suggesting that reading in IELTS focuses on local, lower level processing. Patterns of expeditious reading clearly differed across participants A and B, although the authors indicate that this discrepancy may be due to the search strategies sharing some characteristics which resulted in participants reporting different strategies. Learner awareness of strategy use may also vary across individuals

(proficiency), meaning that any study which seeks to explore these processes must account for individual variation as much as practically possible if variations in processes as they relate to item type are to emerge. Purpura (1999) argues that a lack of evidence of specific strategy usage does not mean that particular strategies were not used. Researchers must account for the possibility that mental processes may go unreported by research participants. The research design of any study may attempt to account for this shortcoming by providing more detailed or innovative stimuli to maximise the opportunity for research participants to verbalise or otherwise report their thought processes.

The TOEFL test developers explicitly identified reading purpose as one of the key extralinguistic features associated with reading language tasks (Jamieson et al, 2000); this is the only situational variable that appears in the TOEFL 2000 ETS monograph (Enright et al, 2000). The test developers believed that increasingly complex reading purpose would influence the cognitive demands placed upon test takers, allowing test users to discriminate between stronger and weaker readers (Jamieson et al, 2008: 71). This continuum of cognitive difficulty (from left to right) conceptualises word identification as the most basic component of academic reading, in contrast to higher-order synthesising of information across texts. Thus the model is to be understood hierarchically, with participants needing to master the skills towards the left of the scale before progressing to those on the right:

Easy cognitive task **Difficult cognitive task**



Reading-to-find-information	Reading for general information	Reading-to-learn	Reading-to-integrate information

Fig. 2.6. Hypothesised implicational scale for reading purposes (Jamieson et al, 2008: 71).

Although the developers created a hierarchy of the cognitive difficulty associated with reading purpose, in the *Reading Framework*, the authors state explicitly that "easy tasks could be designed for reading to learn or reading to integrate [items]… and difficult tasks asking test takers to find discrete information" (2000: 6). Each of these

reading purposes may be associated with text genre (Swales, 1990). The genre will affect the register, which refers to the use of a text and the intentions of the author. Both IELTS and TOEFL texts are intended to have an academic register. IELTS is claimed to contain texts that are of general interest to students at undergraduate or postgraduate level**.** The texts may be written in different styles, for example, narrative, descriptive or discursive/argumentative. At least one text contains detailed logical argument" (IELTS, 2015). Text function also informs the understanding of register in the TOEFL. The TOEFL texts serve the following purposes: expository, argumentative (persuasive) and evaluation; historical/biographical narrative (Enright et al, 2008). The TOEFL is claimed to have "forms that are parallel in terms of register" (Enright et al 2000: 17). In expository texts, the aim is to be informative. Argumentative texts present an opinion by the author and provide evidence to back up this position. Narrative texts contain progressive accounts of events, places or people. Literary texts are avoided due to their cultural specificity.

Text purpose will inform syntactic structure. The role of syntax is important in understanding the construct of reading, but the extent to which this relates to task difficulty depends upon how individual tasks incorporate elements of syntactic complexity. The monitoring column explicitly links test takers' knowledge of grammar and vocabulary and textual features to cognitive processing. Cognitive processing is affected by linguistic variables associated with the input material as well as variables associated with reading purpose. Alderson (2000: 38) provides a taxonomic approach to identifying features of the input that may affect task difficulty and cognitive processing:

- vocabulary,
- syntactic complexity,
- transition markers (cohesion),
- antecedent reference,
- modality (adverbs of attitude),
- amount of text (length), amount of time allowed,

- distances across text when cycle or integration is involved,
- competing linguistic distractors in the text environment,
- cohesion determiners (e.g., “We bought a camera. The lens was cracked.”), grammatical relations as referents (back to subject or back to object), and
- cohesion

Thus, item difficulty is not solely a property of reading purpose, but depends upon the complexity of the cognitive processing that is activated by the item design. This is activated through the features of discourse that the test developers choose to highlight in the text. The extent to which the ability to correctly respond to items depends upon test takers’ ability to parse across complex, multi-clause sentences. This is the heart of cognitive processing and will be discussed in more detail in the next section.

2.5.1.2. Cognitive processing core

Cognitive processing refers to the mental procedures that occurs when test takers perform a task, representing the test taker’s ability to engage with material presented to him or her with the linguistic resources available (language *use*). The conscious strategic management of a task by a test taker (identified in the outside columns) should not be conflated with the cognitive processing of the test taker as they engage with the task (the central core).

In a reading test, the text and item work together to operationalise a desired mental process. It may be possible to infer the nature of the desired cognitive process from these two task features alone. However, this ignores the role of the test taker constructing the meaning generated by this input. Focusing on cognitive processing therefore prioritises the role of the test taker in validation studies (O’Sullivan, 2011). IELTS and TOEFL iBT tests target adult (18+) non-native English speaking populations who have achieved proficiency in their L1 and are then looking to gain admission to a course of academic instruction in their second (or potentially third or fourth) language. L2 readers in the context of this study therefore refers to cognitively-mature

individuals already literate in their first language and are (have been) learning to read in English as a second or foreign language (Koda, 2005). Three basic unique characteristics of this group of learners (Alderson, 2000: 7) are:

(a) L2 readers build their L2 knowledge on their L1 experience (transfer of language learner strategies)
(b) They use knowledge of their L1 to assist L2 acquisition (facilitating structural similarity between L1 and target language)
(c) L2 reading instruction commences prior to gaining L2 oral proficiency (linguistic constraints)

Thus, this group of readers are unique in their reading experiences. Individual learners of English will also be affected by their specific L1 background. IELTS and TOEFL therefore try to capture a range of ability levels in their reading tests.

In a cognitive approach to validity, the question of whether second language reading is specifically a *language* problem or a *reading* problem (Alderson, 1984) gains new prominence, as L2 proficiency is often tested via knowledge of vocabulary or grammar within a text, whereas reading is generally construed as the ability to comprehend the major *ideas* of a text (Koda, 2005: 24). The strength of Khalifa and Weir's model is that the authors have integrated both of these conceptions into the hierarchical central core, where the ability to decode words and comprehend lexical chunks at the clause or sentence level are characterised as lower-level processing, and the ability to comprehend major ideas across sentences is characterised as higher-level processing.

Lower-level processing initially requires lexical decoding, which occurs at orthographic and phonological (sub-word) levels. Orthographic and phonological knowledge contribute to word identification (lexical access). Vocabulary knowledge is crucial to overall text comprehension. Vocabulary knowledge correlates more highly with comprehension than any other factor, including the selection and activation of test taker strategies (Koda, 2005: 49). Efficient visual information processing of individual lexical items is essential for higher-level information processing. Eye movement studies

have demonstrated that each content word received direct visual fixation (Koda, 2005: 30), and have become an increasingly common means of providing validation data in the language testing literature (Bax, 2013; Brunfaut and McCray, 2015; McCray, 2014). These. Nation (2000) identified three major categories of 'word knowledge'; form, meaning and use. Form refers to the morphosyntactic and phonological variations; meaning, the intended denotation and connotation (Anderson and Nagy, 1991) of a specific lexical item and use the grammatical, collocative, pragmatic and illocutionary functions of this term. Hu and Nation (2000) identified the existence of a 'vocabulary threshold'; the authors claimed that adequate comprehension of a text requires knowledge of 98 per cent of the individual lexical items. Core vocabulary (first 2000 words) accounts for approximately 80 per cent of the words in most texts (Nation and Newton, 1997).

Sentence processing requires morphosyntactic structural knowledge to understand how each word contributes to the propositional meaning in a phrase, clause or sentence. At higher processing levels, grammatical and syntactic competencies are insufficient to build mental models of localised areas of the text or build a complete picture of the text. Khalifa and Weir's model is presented as an algorithmic flowchart, implying that each level of processing subsumes the previous levels, with difficulty increasing as one moves up the flow chart. Alderson (2000: 100) refers to these degrees of difficulty as 'processing load'; "the amount of computation relates to the amount of transformation or manipulation necessary to complete a task". Thus, as one moves up the core, the 'processing load' increases.

Koda (2005) also conceives of a cognitive view of reading comprehension in similar terms to Khalifa and Weir. She identifies three clusters – decoding; text-information building; situation-model construction. These are characterised as three competencies: visual information extraction; incremental information integration; text-meaning, moderated by prior knowledge construction (Koda, 2005: 5). The author also noted that in terms of L1 transfer, language learners draw from a number of L1 capabilities in L2 interaction, including knowledge of morphosyntax, phonology, pragmatics, metalinguistic awareness and communicative strategies. L1 linguistic features

influence L2 acquisition and the types of strategies that L2 learners employ. Information-processing procedures vary systematically across languages (Koda, 2005: 15), suggesting there is no universal framework of how L2 learners from different linguistic backgrounds interact with texts. L2 decoding efficiency is at least partially determined by L1-L2 orthographic distance (Koda, 2000).

From the discussion of Khalifa and Weir's model, it is clear that the distinction between lower and higher-level processing requires further elaboration. The boundary lies at the point when participants are required to link ideas across sentences, and knowledge of grammar, syntax and vocabulary is insufficient to discern the writers full meaning. Syntactic structures provide semantic cues for the direction of efficient processing and for what information needs to be retained in working memory. Alderson (2000: 19) notes that "few research studies have isolated significant structural characteristics that deserve to be included as task development variables for testing purposes". Freedle and Kostin (1993) and Freedle (1997) have argued that only a small subset of syntactic variables accounts for difficulty in main idea reading comprehension". The boundary marked by Khalifa and Weir between higher and lower-level processing evokes the discussion in the literature of *cohesion* and *coherence*, which will be elaborated here to provide grounds for understanding this boundary point.

Morgan and Sellner (1980) and Carrell (1982) were concerned with the relationship between cohesion and coherence within a text and whether these were different concepts. For Carrell, coherence is related to *content*, rather than linguistic features of a text, which is identified as *cohesion*. In this sense, textual coherence is a *consequence* of a cohesive text (1982: 482), which is made coherent by the reader's schematic knowledge. Graesser et al (2004) concur with this distinction, arguing that "cohesion is a characteristic of a text, whereas coherence is a characteristic of the reader's mental representation of the text content" (Graesser et al, 2004: 193).

The relationship between cohesion and coherence has been and arguably remains conceptually foggy. Since Carrell (1982), there is agreement that the two concepts are

distinct, but related. In terms of text, cohesion is visible as the product of coherent schema of the writer, which is then observed by the reader who uses the cohesive features of a text to reconstruct the meaning according to their own schematic knowledge. This three-way relationship is visualised in figure 2.7 below:

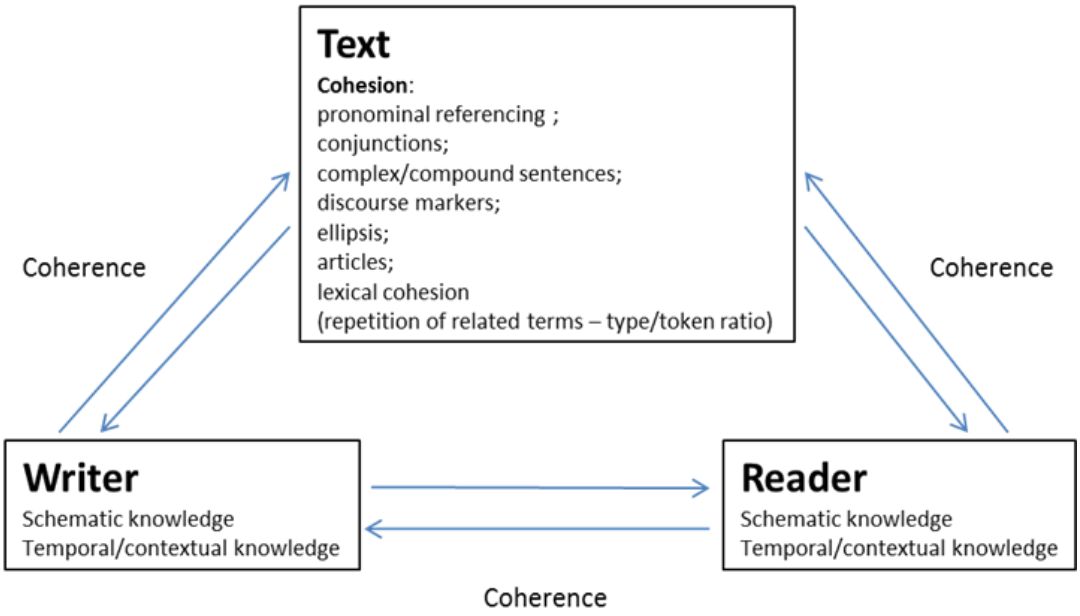


Fig. 2.7. Cohesion and coherence: a three-way schematic map

The diagram is intended to be illustrative and indicative rather than comprehensive or serve as the foundation for the current research programme. Cohesion here refers to observable features of a text in the form of lexis, grammatical decisions and punctuation, which is made coherent as both reader and writer engage iteratively with the text. Coherence is therefore a refined cognitive process. The writer may consistently reflect upon the progress of their argument or narrative, whilst keeping in mind the intended recipient/audience. In turn, upon reading the text, the reader will utilise their own background or topic knowledge (schemata) as a means of interpreting the text. Simultaneously, the reader may formulate an impression of the writer; both their intended argument and/or motivation for creating the text. Coherence located at the base of this model depends upon the congruence of prior experience of writer and reader and their mutual ability to understand relevant lexis for the topic and their mutual understanding of the topic contained within the text.

Cohesion is presented as something *within* a text. Graesser et al argue that a reader will attempt to construct meaning from connections between the constituents of a text, and that this is an inherent part of reading (2004: 194). This is referred to as the *coherence assumption* (Graesser et al, 1994). In addition, the authors argue that cohesion 'gaps' place inferential demands upon the reader, a process which will be affected by schematic knowledge of the topic of the text, and that more advanced language users may in fact benefit from these gaps. The logical corollary of this claim is that cohesive gaps hinder understanding for lower-level language users. This is the distinction between higher and lower-level processing as presented by Khalifa and Weir. This study will be unique in that it will use this model to analyse the cognitive processes in both IELTS and TOEFL to compare them directly.

2.5.1.2.1. Cognitive processing in IELTS

The reading section of IELTS has not undergone a revision since 1995. For a full discussion of the changes that occurred in the IELTS reading section at this time, see Clapham (1996) and Charge and Taylor (1997). It is noteworthy that the current version of IELTS retains many of the features of the features of the original ELTS of the 1980s – specifically comprehension of extended text in reading (IELTS, 2015). Although there is a paucity of research related to the cognitive demands associated with each item type in the IELTS test, Weir et al (2009b) offers the most comprehensive account to date.

Weir et al (2009b) investigated the cognitive processing in the IELTS reading test by using a questionnaire instrument developed in a context validity study comparing the relationship between the IELTS reading test and reading experiences of first-year students at a UK university (2009b). The authors produced an instrument from a mixed-methods approach to identifying how four participants approached a reading task. This produced a strategy-heavy taxonomy that related mainly to the outer (metacognitive and monitoring) columns of Khalifa and Weir's model rather than the

processing core, although some information can be gleaned regarding cognitive processing.

The authors produced Table 2.5 below, in which strategies for specific item types are highlighted if the strategy use recorded for that item type is above a certain threshold value. Differences within item types as well as differences across item types emerged:

Task Type	Section	Text Prev. + mean >.2			Response Strategy + mean > .15											Locating Information + mean > .3				
		PR1	PR2	PR3	ST1	ST2	ST3	ST4	ST5	ST6	ST7	ST8	ST9	ST10	ST11	ST12	LI1	LI2	LI3	LI4	LI5
MCQ	E1.2	+	+		+	+	+						+	+			+				
	E2.1	+	+		+	+	+						+	+				+			
Sent Comp	E3.1	+	+		+	+					+			+			+				
Summ Comp	E1.3	+	+	+	+	+	+						+	+	+	+		+	+		
	F1.2		+		+	+							+	+				+	+		
Heading	F3.1	+	+			+	+	+					+	+				+	+		
	E2.4		+	+				+							+	+			+		
Locate Info	E2.3	+	+			+	+						+	+				+	+		
	F2.1	+	+			+	+	+					+	+				+	+		
Y/N/NG	E1.1		+		+	+	+						+	+				+	+		
	E2.2		+			+	+				+		+	+		+		+	+		
	F1.1		+			+	+	+					+	+				+	+		
	F2.2	+	+			+							+	+				+	+		
Match	E3.2	+	+			+	+						+	+				+			
	F3.2		+	+	+	+							+	+			+	+			

Table 2.5. Text preview, response strategy and locating information by item type (Weir, et al 2009b: 174)

Sequence of reading activities	Strategies for responding to specific items	Information base for the response
PR1 read the text or part of it slowly and carefully **PR2** read the text or part of it quickly and selectively to get a general idea of what it was about **PR3** did not read the text.	**ST1** match words that appeared in the question with exactly the same words in the text **ST2** quickly match words that appeared in the question with similar or related words in the text **ST3** look for parts of the text that the writer indicates to be important **ST4** read key parts of the text such as the introduction and conclusion **ST5** work out the meaning of a difficult word in the question **ST6** work out the meaning of a difficult word in the text **ST7** use my knowledge of vocabulary **ST8** use my knowledge of grammar	**L1** within a single sentence **L2** by putting information together across sentences **L3** by understanding how information in the whole text fits together **L4** without reading the text

	ST9 read the text or part of it slowly and carefully **ST10** read relevant parts of the text again **ST11** use my knowledge of how texts like this are organised **ST12** connect information from the text with knowledge I already have	**L5** could not answer the question

Table 2.6. Strategy taxonomy for IELTS reading test (Weir et al, 2009b)

Information relating to the processing core may be found in the 'information base for the response'. Codes L1 and L2 in table 2.6 mark the boundary between higher and lower level processing as defined in Khalifa and Weir's model. The most commonly reported location for responses was across sentences for all item types, suggesting that IELTS is capable of eliciting higher level processing, and significantly, according to Khalifa and Weir's model, this would indicate that most item types are capable of eliciting higher-level cognitive processing. Item types which did not exhibit this trait were *multiple-choice* and *matching* items. A significant drawback with this study is the self-reported nature of the strategies used, with little direct evidence of engagement with the text and therefore weak corresponding inferences of the level of processing that occurred in relation to specific item types. Purpura (1999) argues that a lack of evidence of specific strategy usage does not mean that particular strategies were not used. Researchers must account for the possibility that mental processes may go unreported by research participants, or conversely, that participants may erroneously report strategies that they did not use. The research design of any study may attempt to account for this shortcoming by providing more detailed or innovative stimuli to maximise the opportunity for research participants to verbalise or otherwise report their thought processes.

2.5.1.2.2. Cognitive processing in TOEFL

For TOEFL, Cohen and Upton (2006) provide the most comprehensive attempt to capture data relating to cognitive processing in the reading component, although similar to Weir et al (2009b), the study focuses more on 'strategies' than 'processes'. The authors' identify two principal strategy types that may be employed by test takers – 'test management strategies' and 'test-wiseness' strategies (a third, 'language

learner strategies', was also identified, but as language learner strategies relate specifically to language instruction rather than assessment, this concept is not discussed here) (Cohen and Upton, 2006; 2012). Test-management strategies are "strategies for responding *meaningfully* to test items and tasks" (Cohen, 2012: 263), that is, explicit or implicit means of responding to the linguistic requirements of the task. In a reading test, this refers to strategies for relating the meaning of the question stem (and if multiple-choice, the options) to the text. Test management strategies are those that operationalize reading skills (Cohen, 2011). For example, a participant identifying a key term in a question stem that leads them to the appropriate portion of a text may be presented as evidence that they s*canned* the text for the relevant lexical item. Test-wiseness is a concept that refers to "using knowledge of testing formats and other peripheral information to answer test items *without going through the expected cognitive processes*" (Cohen, 2011: 264. Emphasis added). Test-wiseness is construct-irrelevant and therefore undermines claims regarding test taker proficiency made from the test scores and observable test taker behaviour.

The authors collected verbal report data from 32 students who were assigned to complete six reading tasks from the LanguEdge Courseware materials which are designed to prepare test takers for the TOEFL iBT. They devised a three-part rubric for reading and test-taking strategies which was used to code participant responses. Due to the different frequency of item types in the instruments, a simple occurrence measure would not accurately report the relative importance of each strategy, as some strategies were reported multiple times in relation to a single item type. The authors developed a simple ratio, whereby the number of occurrences of each strategy was reported in relation to the number of that item type. This provided an impression of the relative importance of that strategy (Cohen and Upton, 2006: 41). The authors also coded video-taped evidence of test-taker behaviour that was not verbally reported, provided such an action could be readily identified. The number of strategies per item type was divided by the number of items of that type to determine the relative importance of each strategy to the test overall. Ad hoc cut-off points were used to determine the frequency of each of the strategies, with four categories being adopted:

Very high (VH) frequency	≥ 1.00
High (H) frequency	≥ 0.50
Moderate (M) frequency	≥ 0.30
Low (L) frequency	≤ 0.29

This was applied to all of the strategies and processes in the authors' rubric in relation to each of the item types:

Strategy	BC-v	BC-pr	BC-ss	BC-f	BC-n/e	I	I-rp	I-it	R2L-ps	R2L-st
R6	H	VH	VH	VH	VH	VH	VH	VH	H	VH
R7	L	M	L	L	L	L	M	H	H	H
R9	M	H	H	VH	H	H	H	H	M	VH
R10	L	L	L	M	L	L	L	L	L	L
R26	L	H	L	L	L	L	L	L	L	L
T1	M	M	M	H	H	H	H	M	M	H
T2	L	L	M	H	M	H	M	M	M	VH
T3	L	L	L	L	L	M	L	L	L	L
T4	L	L	L	L	L	L	L	L	H	H
T5	VH	VH	VH	VH	VH	VH	VH	VH	L	L
T6	L	L	L	L	L	L	M	L	L	L
T8	L	L	L	L	L	L	L	M	L	L
T10	H	M	L	L	L	L	L	L	L	L
T12	H	L	H	H	H	H	H	L	L	L
T13	H	L	L	L	L	L	L	L	L	L
T14	L	L	M	M	M	H	M	L	H	L
T16	H	H	VH	VH	VH	VH	VH	L	VH	VH
T17	L	L	M	L	L	L	M	L	M	H
T19	L	L	M	L	M	L	L	L	H	L
T21	H	L	L	L	L	L	L	L	L	L
T22	L	VH	H	H	H	H	H	H	VH	VH
T24	L	L	M	M	M	M	L	L	L	L
T26	L	L	L	L	L	L	L	M	L	L
T27	M	L	L	L	L	L	L	L	L	L
T28	VH	VH	VH	VH	VH	VH	VH	H	VH	VH

*Frequency rate = no. of occurrences / no. of items of that type. Rates ≥1.0 (marked VH) were classified as 'very high', rates ≥.50 (marked H) were classified as 'high', rates ≥.30 (marked M) were classified as 'moderate', and rates ≤.30 (marked L) were classified as 'low'. Strategies that were used at a low (≤.30) rate across *all* item types were not included in the table.

Table 2.7. Frequency of reported use of reading and test-taking strategies (Cohen and Upton, 2006: 225).

Approaches to reading the passage (R)

R6	Reads a portion of the passage carefully.
R7	Reads a portion of the passage rapidly looking for specific information.
R9	Repeats, paraphrases, or translates words, phrases, or sentences—or summarizes paragraphs/passage—to aid or improve understanding.
R10	Identifies an unknown word or phrase.

Inferences

R26	Verifies the referent of a pronoun.

Test-Management Strategies Coding Rubric (T)

T1	Goes back to the question for clarification: Rereads the question.
T2	Goes back to the question for clarification: Paraphrases (or confirms) the question or task.
T3	Goes back to the question for clarification: Wrestles with the question intent.
T4	Reads the question and considers the options before going back to the passage/portion.
T5	Reads the question and then reads the passage/portion to look for clues to the answer, either before or while considering options.
T6	Predicts or produces own answer after reading the portion of the text referred to by the question.
T8	Predicts or produces own answer after reading questions that require text insertion (I-it types).
T10	Considers the options and checks the vocabulary option in context.
T12	Considers the options and selects preliminary option(s) (lack of certainty indicated).
T13	Considers the options and defines the vocabulary option.
T14	Considers the options and paraphrases the meaning.
T16	Considers the options and postpones consideration of the option.
T17	Considers the options and wrestles with the option meaning.
T19	Reconsiders or double-checks the response.
T21	Selects options through background knowledge.
T22	Selects options through vocabulary, sentence, paragraph, or passage overall meaning (depending on item type).
T23	Selects options through elimination of other option(s) as unreasonable based on background knowledge.
T24	Selects options through elimination of other option(s) as unreasonable based on paragraph/overall passage meaning.
T26	Selects options through their discourse structure.
T27	Discards option(s) based on background knowledge.
T28	Discards option(s) based on vocabulary, sentence, paragraph, or passage overall meaning as well as discourse structure.

Table 2.8. Reading strategies which recorded a rating of 'moderate' or 'higher' in relation to at least one item type (Cohen and Upton, 2006: 220-222).

Cohen and Upton's rubric contains fifty-nine codes. The first twenty-eight are labelled 'reading strategies'. These are split into three sub-categories entitled 'approaches to reading the passage', 'uses of the passage and the main ideas to help in understanding' and 'identification of important information and the discourse structure of the passage'. Of these fifty-nine codes, only twenty-six recorded a rating of 'moderate' or

'higher' in relation to at least one item type. These codes are reproduced in table 2.8 above. The remaining codes recorded sporadic instances only. None of the codes from the processing sub-categories (codes R12 – R25) are included in the table, indicating that Cohen and Upton were only partially successful in capturing processing data.

Of the twenty-six codes in table 2.8, nine (R10, R26, T3, T6, T8, T13, T21, T26 and T27) recorded 'medium' or 'higher' ratings in relation to a single item type. This is particularly noteworthy for code R26 ('verifies the referent of a pronoun'). R26 is identified by the authors as one of only three inferencing strategies included in their rubric, and is only recorded as present in relation to the explicit 'pronoun reference task' (BC-pr), which is one of the most basic forms of inferential reasoning (the item is labelled 'basic comprehension'). No further evidence of inferential reasoning is presented for any of the three item types that specifically claim to measure inferential reasoning (Inference question [I]; Insert text question [I-it]; Rhetorical purpose question [I-rp]). There are several possible reasons. Firstly, that the test items used in the research instruments do not sufficiently encapsulate the meaning of inferential reasoning outlined in the rubric, or that the research design of the Cohen and Upton study was not finely-tuned enough to capture moments of inferential reasoning by the participants. Alternatively, despite coder training, there may not have been enough understanding amongst the coders regarding how to recognise individual moments of inferencing, as these are not explicitly observable; an example of 'method effect'.

An additional example of method effect relates to code R9 ('repeats, paraphrases, or translates words, phrases, or sentences—or summarizes paragraphs/passage—to aid or improve understanding'). This is a strategy that is particular to think-aloud procedures that cannot occur in a live test. This is therefore an artefact of the research process which can actually disguise valuable information regarding how research participants are engaging with the test. Coding a verbalisation 'summarises a paragraph or passage' tells the researcher what the test taker is doing, but in order to gain an understanding of the *cognitive processing* that the test taker is performing, it is necessary to code the summary itself to determine what that summary reveals about how the test taker has understood that particular portion of the text, and whether this

processing has aided the test taker in reaching the correct solution to the item that they are attempting to answer. Finally, a number of the codes are unlikely to be utilised more than once for each item and therefore cannot be rated as 'high' or 'very high' according to the authors' research design. For example, R1, 'plans a goal for the passage' is only likely to be selected once, when the test taker identifies the purpose of the question by reading the question stem. Returning to the question for clarification is coded differently as a test management strategy (T1-3).

It is notable that the principle studies cited here as aiming to provide evidence of cognitive processing for both IELTS (Weir et al, 2009a, 2009b) and TOEFL (Cohen and Upton, 2006) have both been predicated on verbal evidence provided by test takers. This section has highlighted some weaknesses with the outcomes of the studies in terms of identifying cognitive processes in particular, which require further critical engagement to determine the utility of this approach for the current study. The next section therefore explicitly examines the strengths and weaknesses of verbal protocol analysis in identifying cognitive processes in L2 reading.

2.5.2. The use of verbal protocol analysis to study cognitive processing in L2 reading

Verbal protocol analysis is a data collection procedure in which participants verbalise their thought processes that they used to complete a task. Verbalisations are recorded and transcribed for analysis. Researchers may adopt a grounded approach or analyse verbalisations according to preconceived research framework. Protocol analysis is a well-established research method for investigating L2 reading (Bereiter and Bird, 1985; Block, 1986; Anderson, 1991; Anderson et. al., 1991). Procedural verbalisations during which participants are not required to justify their comments or explicitly reflect upon them are termed *non-metalinguistic*. Verbalisations in which participants are explicitly asked to reflect upon and explain their reasoning are termed *metalinguistic*. Bowles (2010) reinterprets these terms as 'non-metacognitive' and 'metacognitive' respectively to apply more readily to both verbal and non-verbal tasks (2010: 13) in second language research.

Verbal reports that occur simultaneously as the task is completed are known as *concurrent* verbal reports, whilst those that are elaborated after completion of the task are termed *retrospective* verbal reports (Bowles, 2010: 1). The argument implicit in the deployment of verbal protocols is that this procedure necessarily provides an accurate representation of participants' thought processes. There are, however, four caveats to that argument which potentially undermine verbal protocols as a valid and reliable research methodology for the current study. These are veridicality (truthfulness), automaticity, reactivity and the language of utterance (L1/L2).

'Veridicality' refers to the truthfulness of utterances. That is, to what extent do the verbalisations truly reflect the thought processes of the test taker. Ericsson and Simon (1980: 110) note that online metacognitive verbal reports may alter cognitive processing under scrutiny in addition to requiring extra time to complete the task (latency) (1980: 52-54). If participants are expected to complete two tasks simultaneously, this imposes an additional cognitive load which will either affect performance, or alter the processing sufficiently that verbalisations no longer reflect those that test takers would otherwise undergo during task completion. They also argue that verbal reports will differ depending on the situation in which they were collected (1980: 221) and the level of detail that is reported by participants (metacognitive or non-metacognitive) (1980: 90-91) will vary as a result.

'Automaticity' is the effortless, non-deliberate retrieval of specific activation patterns previously stored in memory (Koda, 2005: 18). Specific strategies that have *become* unconscious are described as automatized or *subconscious* (Faerch and Kasper, 1987). Automaticity is beyond the control of individual test takers, whereas strategy usage is regarded as "deliberate and purposeful" (Cohen and Upton, 2006: 3). This is a problem as participants are unlikely to be able to explain actions that have become automatized. Bereiter and Bird (1985: 132) argue that "rapid and automatized… [cognitive] protocols taken during reading are relatively impoverished compared to those that emerge in more deliberative activities". If a 'strategy' is described as something consciously deployed, then automatized procedures may no longer be

classified as strategies (Cohen, 1998; Ellis, 1994), although Phakiti (2008) notes that some test takers may underline or highlight specific areas of a text without realising that they do so, in contrast to others who consciously adopt such a strategy. Such actions may be captured on video for playback to the research participant.

'Reactivity' refers to the concept of verbal protocols interfering with the cognitive processing of research participants (Bowles, 2010: 14). Ellis (2001: 37) argues that 'dual-processing' inherent in think-aloud verbal reports is problematic for less proficient L2 learners. Therefore, proficiency of participants represents a moderator variable that must be accounted for in verbal protocol studies. Leow and Morgan-Short (2004) examined reactivity to non-metacognitive think-aloud reports with 77 beginner learners of Spanish in a reading comprehension task and did not find a statistically significant difference between the think-aloud and control group, suggesting that reactivity was not a crucial variable in performance. Bowles and Leow (2005) examined the reactivity of metacognitive and non-metacognitive think-alouds on a higher-proficiency Spanish comprehension exercise containing the pluperfect subjunctive with 45 fifth-semester Spanish learners. Compared to the control group, non-metacognitive verbalization once again did not affect performance on the task (2010: 68).

Polio and Wang (2010, cited in Bowles, 2010) compared the text comprehension of a non-metacognitive think-aloud group to a silent control group, replicating the Leow and Morgan-Short (2004) study with more advanced participants. A significant difference emerged in comprehension between the two groups (n = 15 for each group). Participants in the think-aloud group comprehended the text worse than the silent group. This contrasts with the findings of Berardi-Coletta et al (1995), who found that a metacognitive approach to think-aloud tasks could actually improve task performance by participants at a statistically significant level. One hypothesis could be that think-aloud tasks preserves commitment to goal orientation. Veridicality and reactivity allow critics to question the completeness of utterances and raises doubts as to the validity of inferences made from verbal data, and whether such a data collection tool hinders replicability of findings (reliability). Memory lapses between task

completion and a retrospective think-aloud session may hinder the accuracy of process or strategy identification.

In verbal protocol analysis, participants may be asked to verbalise in English or their L1. Verbalisation in L1 has the advantage of allowing participants to focus on the content of their verbalisations rather than translating the thoughts before uttering them. In the event of L1 verbalisation, verbalisations must be translated by an independent translator, although this will add time, cost and an additional layer of interpretation to the research. If the aim of the verbal protocol analysis is to ask participants to produce non-metacognitive verbalisations, then participants may be able to use simpler English language verbalisations. L2 verbalisation also determines the proficiency level of participants who may be recruited for the study. Lower ability participants will not be able to adequately verbalise their thought processes. Nyhus (1994) argues that L2 participants believe there to be a threshold above which they are able to effectively participate in L2 think-aloud research. Below this threshold, they feel that their language proficiency is a hindrance to their participation. Unfortunately, Nyhus (1994) did not include CEFR-calibrated test scores, suggesting this threshold of participation remains under-explored and could form the basis for any future research into verbal protocol analysis conducted in participants' L2. An inability to verbalise their thought processes will result in incomplete verbalisations or questionable veridicality, from which coders will not be able to derive meaningful statements regarding participants' strategy use or cognitive processing. Therefore, objective evidence of participants' proficiency in the form of test scores (for both reading *and* speaking) must be collected in order to claim participants are able to take part in the research. Tense difficulties can also cause confusion regarding whether participants are truly referring to the past or the present. Mackey, et al (2000) determined that in a stimulated recall in which Italian ESL learners were required to verbalise in English during the recall session, they addressed significantly less of the stimulus material than English L1 speakers.

In verbal protocol analysis, each of these issues must be addressed in order to satisfy the requirements of veridicality that the participants' utterances are accurate reflections of their thought processes. From this, claims about the cognitive processing

required in successful task completion can be obtained. The final section in the literature review brings together the sections on ECD, RE and cognitive processing in reading in order to develop relevant research questions.

2.6. Development of the research questions

The literature review has covered multiple areas that are pertinent to the current study. The study proposes to demonstrate the utility of the framework RE, which will be illustrated through comparative analysis of the reading components of the IELTS and TOEFL tests. Part 1 introduced the concept of reverse engineering and the theoretical development of this concept, which was enhanced by reviewing relevant literature from computer science, and demonstrated how a framework of RE could be built upon the principles of evidence-centred design. This section also considered how the concept of validity in relation to ECD would impact on validation in the subsequent RE framework. Part 2 specifically considered the literature regarding cognitive processing related to reading, with particular emphasis on the framework of Khalifa and Weir (2009). The final part then considered a key methodology (verbal protocol analysis) regarding the elicitation of evidence for cognitive processing in relation to the reading tests.

The study is concerned with the development of a theoretical framework of RE. This framework has been developed in the literature review. The study is concerned with testing this framework. Therefore, the overall guiding research question is

> *'Can an evidence-centred framework of reverse engineering be used to develop representative cognitive test specifications for the purposes of test comparability?'*

To test the framework, this study will focus on a demanding critical parallel research project. This will be achieved by a comparability study of the IELTS and TOEFL iBT reading sections. Thus, the study will also address longstanding issues of test comparability between IELTS and TOEFL. The study focuses on identifying the

respective test developers' conception of the construct to produce a 'cognitive specification' for each of the item types in IELTS and TOEFL. To identify the relevant cognitive processes, the study has developed a theoretical framework for RE based on Robert Mislevy's evidence-centred design (ECD) (see figure 2.4). The study is a critical parallel analysis, meaning that the central core of the framework, composed of the three components of ECD which inform the research design are the evidence model, the student model and the task model. Table 2.7 below identifies the relevant research questions associated with each ECD component.

Principal research question:	**Can an evidence-centred framework of reverse engineering be used to develop representative cognitive test specifications for the purposes of test comparability?**
ECD Component	**Research questions for a *critical parallel* reverse engineering agenda**
Evidence Model	**Research question 1:** *What observable test-taking strategies do test takers use when completing IELTS and TOEFL reading tests? Are there any differences in how participants respond to IELTS and TOEFL tests?*
Student Model	**Research question 2:** *Which cognitive reading processes do test takers use when they complete IELTS and TOEFL iBT reading sections?*
	Research question 3: *Do these processes reveal differences in IELTS and TOEFL iBT test developers' understanding of the construct of interest?*
Task Models	**Research question 4:** *Are cognitive processes associated with specific item types? Do individual item types target specific processes or do they elicit a range of processes?*

Table 2.7. Research questions

Research question 1 relates to the evidence model. As outlined in section 2.4, the evidence model details the evidence required to make inferences from observable variables (work products) to the construct of interest. Placing the test taker at the centre of the RE effort means that observable evidence takes the form of test taker interaction with the tests. The work products are the test taker responses to the items. Research question 1 aims to identify the task completion strategies that participants use to complete the IELTS and TOEFL tests.

Research questions 2 and 3 relate to the student model. The student model describes the construct in terms of what is being measured. The construct is interpreted in this study as the cognitive processes that the test takers employ in order to complete test tasks. By definition, these are unobservable, and must be inferred from observable interaction between the test takers and the test tasks, in the forms of their responses to individual items and their observable test taking strategies. Research question 2 specifically aims to identify those cognitive processes. Research question 3 then asks whether there are any differences between how participants complete the IELTS and TOEFL tests.

Research question 4 relates to the task models. In critical parallel RE, task models consider whether specific tasks are associated with specific claims made about successful test takers. Whereas research questions 2 and 3 consider the data pertaining to the tests as a whole, research question 4 asks whether specific cognitive processes can be identified with specific item types, or whether the test developers intend that individual items target a range of cognitive processes. Thus, the methodology will have to be sufficiently finely-grained to collect data at the item level.

The next section (methodology) provides in-depth exploration of the issues associated with the research questions and how the data will be collected. The chapter will begin by defining the key terms that will be used in the study, based on the discussion of these terms in the literature review. The chapter will outline the research design and how this addresses the advantages and disadvantages of the selected methodology based on its underlying epistemology. The study details three initial studies that were conducted to assist the selection of an appropriate research methodology, and an additional pilot study to explore these issues and assess whether the approach could deliver the data needed to address the four research questions. The chapter then explicitly outlines the finalised research design and provides details of the participants and ethical procedures. The methodology then details the means by which the data will be analysed.

Chapter 3. Methodology

3.1. Introduction

This section provides in-depth exploration of how data to address the research questions will be collected. The chapter begins by defining the key terms used in the research questions and how they are to be understood in the context of this study. The chapter will then outline the research design on the basis of the RE framework and introduces the procedure that the critical parallel study will follow.

The chapter will then address the selected methodology and how it will provide relevant data for the research questions. This chapter also introduces the epistemological assumptions of the methodology and how these link to the interpretive argument that underpins the RE framework. This chapter details three studies that were conducted to assist the selection of an appropriate research methodology, and an additional pilot study to determine whether the method could deliver the data needed to address the four research questions. The chapter also provides details of the participants and ethical procedures followed. The methodology also details how data will be analysed.

The new framework of an evidence-centred approach to RE resulted in the development of the following inter-related research questions, predicated on the respective elements of Mislevy's (2003) conceptual assessment framework (CAF) as outlined in the literature review:

1.) What observable test-taking strategies do test takers use when completing IELTS and TOEFL reading tests? Are there any differences in how participants respond to IELTS and TOEFL tests?
2.) Which cognitive reading processes do test takers use when they complete IELTS and TOEFL iBT reading sections?
3.) Do these processes reveal differences in IELTS and TOEFL iBT test developers' understanding of the construct of interest?
4.) Are cognitive processes associated with specific item types? Do individual item types target specific processes or do they elicit a range of processes?

These research questions and the data gathered to address them will be used to address the overall research question on which the study is predicated:

> *Can an evidence-centred framework of reverse engineering be used to develop representative cognitive test specifications for the purposes of test comparability?*

3.2. Defining the key terms used in the research questions

The key terms used in the research questions are 'test-taking strategies' and 'cognitive processes'. These terms were selected based upon the literature reviewed in the previous chapter. Clarification of key terms is essential for designing an appropriate methodology that addresses the research questions. Grabe (2000) and Afflerbach et al (2008) point to a lack of consistency regarding the use of terms such as 'strategies' and 'processes' which may result in terms being erroneously used interchangeably. The terms are therefore deliberately used in separate research questions relating to different models of the CAF. 'Strategies' are defined as procedures that are used to respond meaningfully to test tasks, as defined by Cohen (2012: 263). 'Observable' in the research question refers to those strategies that the participants themselves are able to report, and others that they may not be consciously aware of, but which may be captured by some other means.

For research questions 2 – 4, the definition of 'cognitive processing' refers to the mental procedures that occurs when test takers perform a task; the definition offered by Khalifa and Weir (2009). Cognitive processing therefore refers to the linguistic resources that enable test takers to establish meaning between the task and the input text. For the purposes of clarity in this study, 'strategies' refer to the conscious decisions made about goal-oriented actions, with 'cognitive process' specifically referring to the meaning making that the test taker derives from the test task, text and their own linguistic resources. Research question 2 aims to identify individual processes. This research question requires some reliable means of identification.

Research question 3 aims to compare IELTS and TOEFL directly, so therefore requires some means of quantification. Research question 4 requires that the methodology be sufficiently finely-grained to identify whether individual processes are linked to specific items, or whether individual items target a range of processes.

3.3. Overview of the research design

With a commitment to collecting highly detailed and context-specific data, this thesis adopts *stimulated-recall interviews* (SRI) as the principal method of data collection for all four research questions. SRI are an introspective and retrospective method of verbal protocol analysis designed to "elicit verbal data about the thought processes involved in carrying out a task or activity" after the event (Gass and Mackey, 2000: 1). Data in the form of observable stimuli was interpreted as the means of identifying strategic behaviour to address research question 1. These stimuli were presented to participants to elicit verbal data which was analysed to make claims about cognitive processing to address research questions 2 – 4.

The methodology was chosen based upon the review of literature which identified verbal protocol analysis as the principal data gathering methodology regarding metacognitive strategies and cognitive processes in L2 reading. Three initial studies, reported in this chapter, aimed to identifying an approach to verbal protocol analysis which would be most useful in collecting data for the four research questions. The next section discusses the philosophical perspective which underlies SRI and how these inform the study, followed by an account of three initial studies that were conducted to inform the research design and how the findings led to decisions being made about the methodology. The outcomes of the three studies are used to identify key methodological considerations of SRI which are explicitly identified in order to inform the research design. The finalised research design informed the pilot study, which was conducted in order to evaluate the quality of the data collected.

3.3.1. Philosophical perspectives of SRI

In this study, the aim is to use an RE framework and SRI methodology to identify which of the elements contained within Khalifa and Weir's model can be inferred from participant verbalisations. Theoretically, this serves two purposes. First, demonstration that a principled approach RE can be achieved in small-scale studies by locally-based stakeholders and that the outcomes of this research can be used to make claims about the nature of the target domain as envisioned by the test developers. As a subsidiary claim, the research will also provide empirical evidence for the different levels of Khalifa and Weir's (2009) cognitive processing model of reading.

The efficacy of SRI as a research methodology hinges upon two key underlying assumptions – that participants are capable of remembering and identifying the unobservable processes that they utilised to complete a task, and that they have the capacity to verbalise those processes. These assumptions may be characterised as *procedural* and *declarative* knowledge respectively (Gass and Mackey, 2000: 21). Declarative knowledge is directly available to participants and therefore researchers through introspection, whereas "procedural knowledge is comprised of the cognitive and interactional processes involved in the reception, production and acquisition of knowledge… [which] is considered automatic and inaccessible via introspection" (ibid.: 21). Thus the introspective technique requires *inferential reasoning* from the declarative statements regarding participants' actions to the cognitive processing that underpins those verbalisations and the actions to which they refer. A model of inferential reasoning is therefore necessary to reliably identify specific processes from specific verbalisations to make claims regarding the type of cognitive processing occurring in the mind of test takers.

Toulmin's (2003) model of argumentation provides the epistemological grounds for making claims regarding cognitive processing from participant verbalisations which forms part of the ECD approach to test validation. Toulmin's logic model was used to consider the implications of adopting the SRI methodology developed in the initial and pilot studies to the type of data that the study required (figure 3.1). This process of logical argument encompasses the methodological issues associated with SRI and how

they support or detract from a claim of specific cognitive processes made from participant verbalisations:

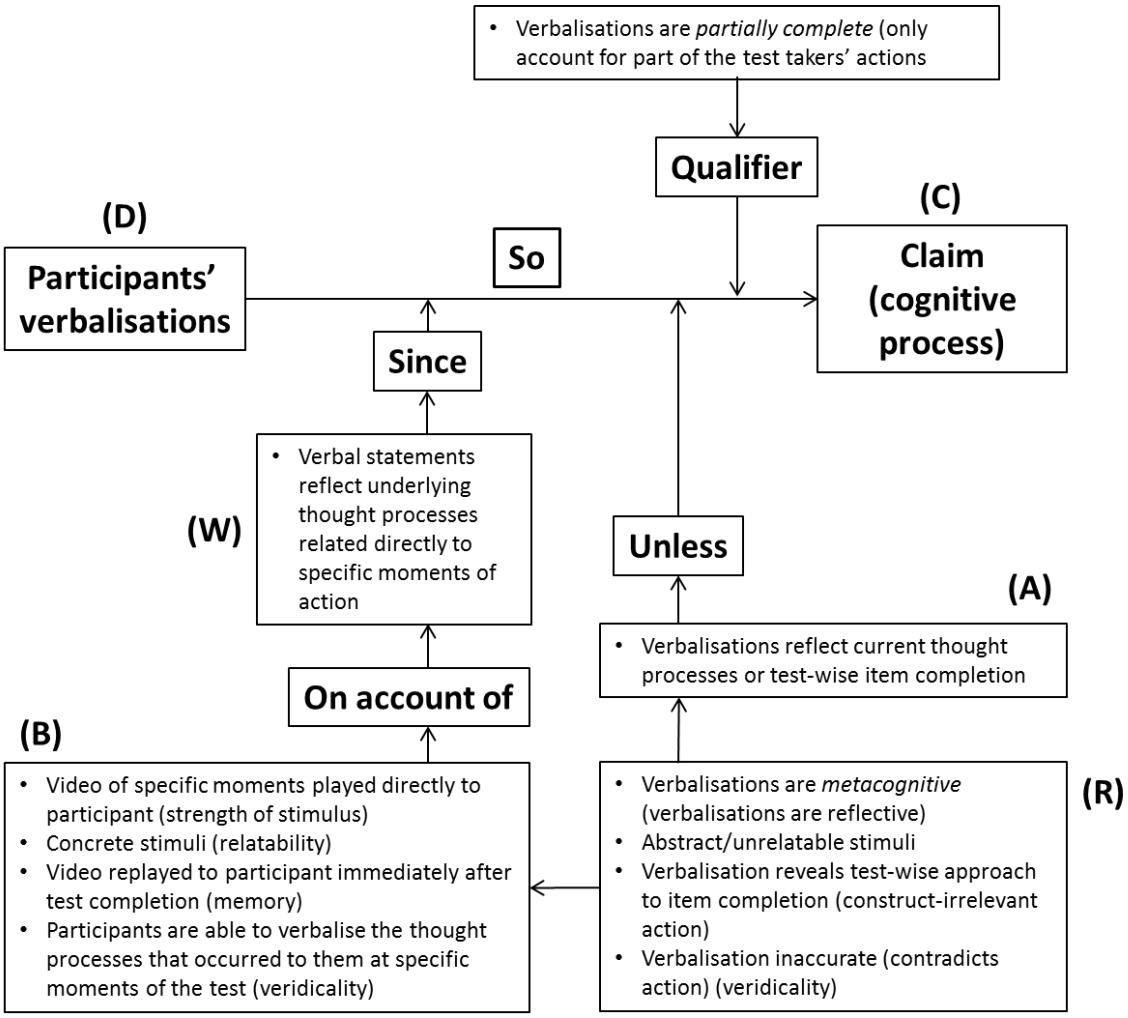


Figure 3.1. Toulmin diagram of logical argumentation to relate data to claims

The figure encompasses the inferential argument used to make claims regarding participants' cognitive processes on the basis of participant verbalisations in SRI (the 'so' process arrow, linking data with claims). Other boxes contain characteristics of the verbalisations and the conditions of the SRI that will impact claims (C) about individual test takers, items and tests on the basis of the collected data (D). The *warrant* (W) states that the verbalisations relate directly to thought processes at specific moments of action. The alternative explanation (A) argues that the verbalisations made in the SRI do not relate to the thought processes during test completion, but are a

metacognitive record of their thoughts at the moment they utter them about the processes as they occurred. A second alternative explanation is that the verbalisations reflect the thought processes at a given moment in test completion, but these verbalisations reflect test-wiseness rather than a cognitive process specifically related to item completion. The backing (B) and the rebuttals (R) provide evidence for the warrant and alternative explanations respectively. The content of these boxes relate to the conditions of the SRI, and reference the *ability* of the participants to verbalise meaningfully (in English), the *relatability* of the stimuli to act as effective recall material and the *conditions* of the SRI in terms of timing in relation to the target performance.

The study is predicated on making claims about *tests* rather than test takers. Test developers provide instruments for stakeholders to make claims about individual test takers. The strength of these claims depends upon the consistency of measurement across test forms. Therefore, for test developer claims to withstand scrutiny from stakeholders, each form of their test instruments must be a representative example of the construct as understood by the developers. Generalisable claims in the present study do not concern people, but *ideas* (related to the processing inherent in the test instruments). This is a form of *analytic* generalisation (Yin, 2010).

Following Yin (2010), the present study adopts the following elements of a research programme undertaken for the purpose of analytic generalisation. Claims made on the basis of participant verbalisations are founded on a form of logical argumentation used by Kane as part of an evidence-centred approach to validating test claims. Toulmin's (2003) model of argumentation accounts for the possibility of rebuttals and alternative arguments being put forward to account for the data. The research design will include multiple participant verbalisations for each item in both tests, strengthening claims about the cognitive processes required for particular items in which the participants exhibit the same level of processing. Outcomes in the form of claims regarding cognitive processes are based on the cognitive processing core identified by Khalifa and Weir, not produced from grounded argument. The findings will be used to evaluate which of the elements of Khalifa and Weir's model are relevant for the IELTS

and TOEFL tests and the extent to which a level-based understanding of reading processes is justified by the data. The study will also determine the efficacy of a small-scale, principled approach to RE and how this may be used in studying other skills as part of an ongoing RE project.

3.3.2. Three initial studies to determine the approach to verbal protocol analysis

Three initial studies were conducted in order to establish the formal procedure for the verbal protocol analysis. One participant was selected for each of the three studies. These participants were all Chinese, in order to control for L1 variation. An IELTS test was selected at random for the initial studies. Section 3.5.2 details the principled methods by which test instruments were selected for the main study.

3.3.2.1. Initial study 1

The participant who undertook the first study is a female PhD student at the School of Education at the University of Leicester. Her most recent IELTS score is 6.0. She undertook the IELTS test in July 2009, prior to commencing her doctoral studies. The required entrance score to undertake the programme was 6.5, which necessitated her participating in a four-week EAP pre-sessional at a University English language teaching centre.

The participant was presented with an IELTS text with thirteen associated items representing two item types (matching headings and author opinions). A retrospective verbalisation procedure was agreed upon, during which the participant would speak freely without input from the researcher. Upon completion, the participant activated a digital voice recorder and recorded her thoughts about the testing process. Excerpts of her verbalisations are outlined below:

Transcription

*"I found it was not pleasant. Firstly**, I have no time control**. I did not bring my watch so I don't know. I am told I need 20 minutes to finish the test, but **I didn't know how to manage the time well**, so that was frustrating. Secondly, I have done this kind of reading before, so I know that when I do the readings, if I want to find the correct answers, I [have] **got to read all the information as soon as possible and as clear as possible**, so I first looked at the **key words** and **each paragraph, there are key words linked to the headings** and that I should **circle them**. This could maybe help me to find the correct answer."*

*"So after, **if I have problems with for example B, I go straight to C**. B I crossed first, at the start of my reading, I crossed B **and the answer from the heading as well**, so I don't need to bother with this as a strategy for the time control. I found this reading quite… the information isn't familiar so it's a bit difficult for me to read**, even though I understand [the words] I don't need to understand all the information**. Even though, I still couldn't find the answer. I used scanning, and **I did not really read all the texts word by word, just found the key words**, and then for question 14 to 19, I **used strategies like crossover [crossing out]**. I just crossed one… it's kind of **avoiding me to read the same thing again**, or to read the same headings again. For the questions 20 to 26, in the same way, I looked at the **[unclear]** like **the key words** and here I looked at, 20 for example, I look at FAA, and then I **quickly target** FAA from the reading text, so I **located the place and I found the rest of the verse**, for example the 'jet engine'; **if I don't find this, or if I find some of the information is not similar to this, I could say it's 'false'**. So I used the same strategies for the rest of the questions. That's my experience with the reading."*

Table 3.1. Initial Study 1 transcription

The participant recorded a score of 6/13. A stimulated-recall approach was adopted to minimise reactivity. Nonetheless, the participant still achieved a low score, likely due to the low-stakes testing situation. In order to make claims regarding the required cognitive processing to complete individual items (and therefore the tests), higher-scoring performances are needed. Asking two test takers to complete the same instrument would increase the probability of observing a correct response for each item, and thus the ability to make claims about that item. It also provides a point of triangulation. Response processes can be compared to identify differences and similarities regarding item completion processes.

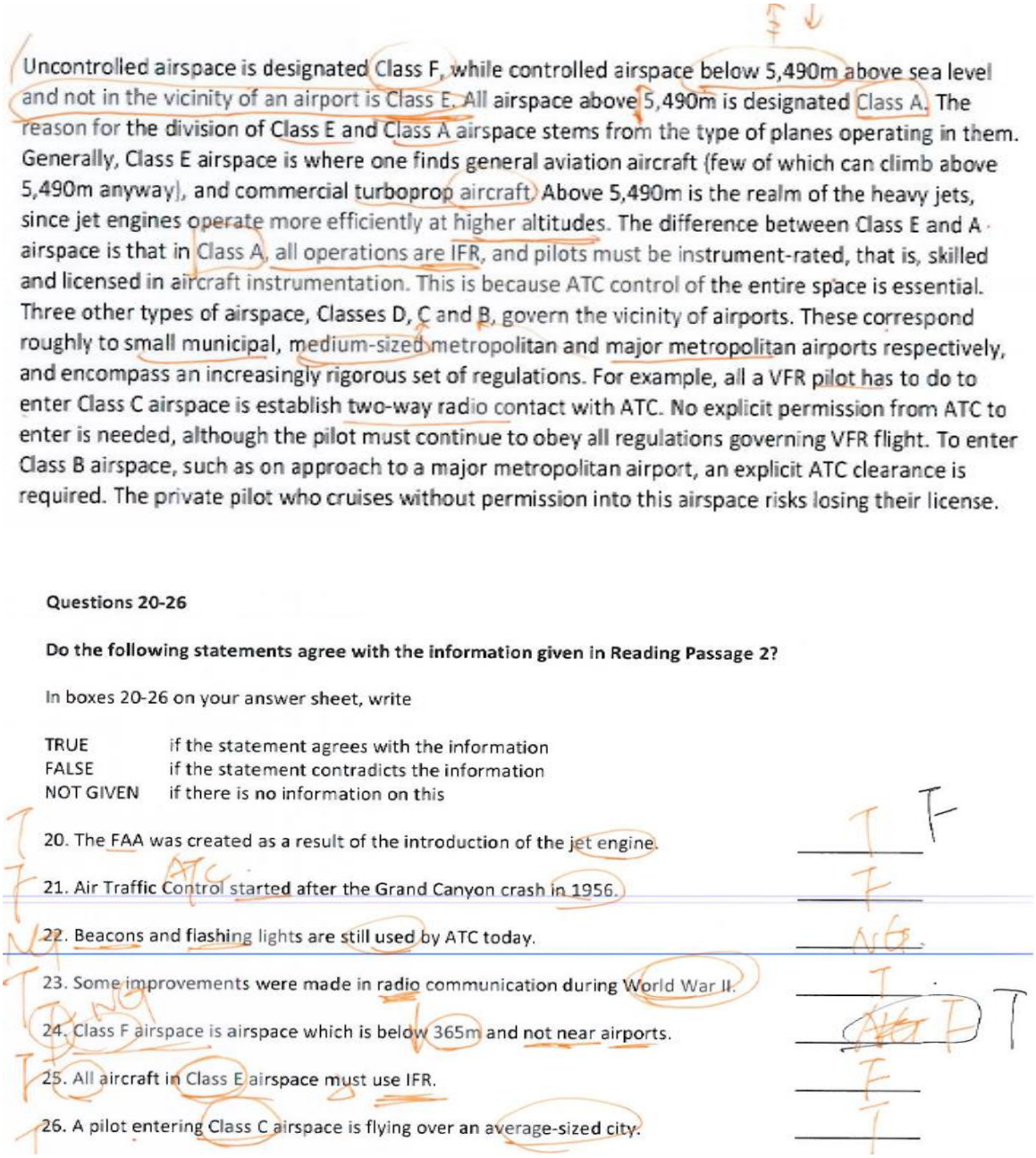
The participant was especially aware of the importance of time management and the subsequent impact this has on how to proceed with the test. She did not have a watch and was unable to manage her time effectively. Taking a test under exam conditions replicates the conditions for a high-stakes test, although possibly reduces the likelihood of observing correct responses. As this study requires usable data, facilitating participant verbalisations is of greater concern than ecological validity.

The awareness of the time constrictions immediately influences the subsequent strategies that the participant applies to test completion. She considered the most useful approach to item completion to be key word identification in both item stems and in the text (expeditious reading), thereby allowing her to quickly navigate to the relevant portions of the text. Particular words, phrases or initialisms that stand out in the question stems may be found quickly and efficiently in the text. Evidence suggests that the participant followed this strategy and then read around these words to establish propositional meaning in order to answer those items (careful reading). She then marks (circles) key words. Focusing on key words rather than overall impression of the meaning of the text possibly prevented the participant from forming building a mental model or establishing text-level understanding, potentially explaining why she recorded a score of 6/13.

An observable test management strategy that the participant employed is completing those items that appear easier first. Thus more challenging items do not monopolise the test taker's time. A very explicit test management strategy was crossing a heading from the list of options once it had been matched with a respective paragraph to avoid considering this heading for future items. The weakness of such an approach is requiring the participant to be extremely confident about that match. There is no evidence from the transcript of the participant revisiting any of the questions to reconsider her choices once made, possibly due to the time constraints. A problematic feature of the verbalisations is the tendency to provide an explanation for *item types*, rather than individual items. Verbalisations related to item type assume that identical cognitive processes are elicited for each example of a single item type. This ignores the understanding of the text that the participant develops as she answers individual items, making it more likely that she would adopt fewer search strategies once she is aware of the location of specific information.

3.3.2.2. Initial Studies 2 and 3

For the second and third studies, two test takers were assigned the same instrument as the first study and the same allotted time (20 minutes). A more controlled SRI followed immediately after the test was completed, in which the participants were probed as to how they completed each item. Participant responses and the annotated text served as the stimuli. Examples are reproduced here:



Uncontrolled airspace is designated Class F, while controlled airspace below 5,490m above sea level and not in the vicinity of an airport is Class E. All airspace above 5,490m is designated Class A. The reason for the division of Class E and Class A airspace stems from the type of planes operating in them. Generally, Class E airspace is where one finds general aviation aircraft (few of which can climb above 5,490m anyway), and commercial turboprop aircraft. Above 5,490m is the realm of the heavy jets, since jet engines operate more efficiently at higher altitudes. The difference between Class E and A airspace is that in Class A, all operations are IFR, and pilots must be instrument-rated, that is, skilled and licensed in aircraft instrumentation. This is because ATC control of the entire space is essential. Three other types of airspace, Classes D, C and B, govern the vicinity of airports. These correspond roughly to small municipal, medium-sized metropolitan and major metropolitan airports respectively, and encompass an increasingly rigorous set of regulations. For example, all a VFR pilot has to do to enter Class C airspace is establish two-way radio contact with ATC. No explicit permission from ATC to enter is needed, although the pilot must continue to obey all regulations governing VFR flight. To enter Class B airspace, such as on approach to a major metropolitan airport, an explicit ATC clearance is required. The private pilot who cruises without permission into this airspace risks losing their license.

Questions 20-26

Do the following statements agree with the information given in Reading Passage 2?

In boxes 20-26 on your answer sheet, write

TRUE	if the statement agrees with the information
FALSE	if the statement contradicts the information
NOT GIVEN	if there is no information on this

20. The FAA was created as a result of the introduction of the jet engine. _____
21. Air Traffic Control started after the Grand Canyon crash in 1956. _____
22. Beacons and flashing lights are still used by ATC today. _____
23. Some improvements were made in radio communication during World War II. _____
24. Class F airspace is airspace which is below 365m and not near airports. _____
25. All aircraft in Class E airspace must use IFR. _____
26. A pilot entering Class C airspace is flying over an average-sized city. _____

Figure 3.2. Stimuli: Annotated text and question stems from initial study 3

The participants were asked for their most recent IELTS score (or other English language test score that they may have been awarded), the reading score that they

obtained, the date of administration, whether or not they undertook an EAP course prior to commencing their academic programme (and its duration) and their academic pathway. This information is included in table 3.2 below:

Participant 2 (Initial Study 2)	Participant 3 (Initial Study 3)
• Overall IELTS score of 6.5 (test taken July, 2012) • Reading score of 6.0 • Participant is currently studying on a Medical Physiology foundation course at the University of Leicester. The foundation course has a focus on English for medical purposes. • The participant has not attended any courses at the ELTU.	• Overall IELTS score of 5.5 (test taken April, 2012) • No reading score available (the participant was unable to recall). • Participant arrived in Leicester in October 2012. They studied three courses at the ELTU (Courses, B, C and D). • The participant is intending to undertake a Master's degree in Financial/Mathematical modelling in the Department of Economics.

Table 3.2. Participant details for initial studies

Areas of interest to the second and third initial studies were the duration of the test/interview and the sufficiency of the evidence for making claims related to the strategies and processes employed by the participants. Participant responses to the 13 items are outlined below. It is noteworthy that participant 2 holds an IELTS score of 6.5, in contrast to the 5.5 score of participant 3, yet the latter recorded a score of 8/13 versus 5/13 for the former. In mitigation, the IELTS score for participant 3 precedes the research by more than 12 months, a time period during which her proficiency likely improved substantially. Additionally, a short instrument (only one-third of a complete IELTS reading test) is insufficient to draw inferences or make claims regarding the overall reading proficiency of the participants. This data was collected as a form of evidence that participants were able to participate in the research.

Participant 2 responses:			**Participant 3 responses**			**Correct responses:**
14.	v	x	14.	i	x	14. ii
15.	ix	x	15.	ii	x	15. iii
16.	viii	x	16.	v	✓	16. v
17.	iv	✓	17.	iv	✓	17. iv
18.	vi	x	18.	iii	x	18. viii
19.	vii	✓	19.	vii	✓	19. vii
20.	T	x	20.	T	x	20. F
21.	T	x	21.	F	✓	21. F
22.	NG	✓	22.	NG	✓	22. NG
23.	T	✓	23.	T	✓	23. T
24.	F	x	24.	F	x	24. T
25.	NG	x	25.	F	✓	25. F

26. T ✓	26. T ✓	26. T

Table 3.3. Participant Responses for initial studies 2 and 3

Excerpts of participant verbalisations are included in table 3.4 below. The column on the right contains inferential statements made by the researcher related to the strategy usage by the participants and level of cognitive processing related to Khalifa and Weir's model (2009). Verbalisations indicative of specific levels of Khalifa and Weir's model emerge. Verbal evidence of word matching may be characterised as 'lexical access' if the participant demonstrates that the action was driven by understanding of that word rather than simple orthographic matching. If participants demonstrate that the sentence structure (e.g. referencing function rather than content words) was important in their actions, then this can be labelled 'syntactic parsing'. Verbal statements in which the participant demonstrates understanding of the meaning of a specific sentence can be labelled 'establishing propositional meaning'. However, there was no evidence in these initial studies of higher-level processing:

Initial Study 2		**Comments**
Interviewer:	*So, we'll go through it item by item. So, question 14, paragraph A, you chose title 1. Can you explain why you chose title 1?*	The student tried **expeditious reading** due to the time constraint. However, she was unable to deduce the meanings of key unknown words, e.g. 'establishment' **[lexical access]**, which could have helped her to choose the best heading. Instead, she resorted to a remedial strategy (**matching a heading** that contains the same word as in the paragraph **[lexical access]**).
Participant:	*Yeah, because when I read this, I first read the titles, and then went to find something… this way, paragraph A also said 'what is the FAA,' so generally speaking, it just introduced [it], so I chose one.*	
Interviewer:	*So, it's matching 'FAA' in the text to the title…*	
Participant:	*Yeah, because it says the FAA regulations, so it [matched]…*	
Interviewer:	*OK. Number 15, paragraph C, you chose title 2.*	This verbalization reveals a failure of **expeditious reading.** The student read only topic sentences (at the beginning of a paragraph) for exam purposes under timed conditions, hence missing important information that comes in the rest of the paragraph. She matches key information **[lexical access]** but no higher level of processing is evident.
Participant:	*Yeah because in this paragraph it says 'new development' so maybe it suggested improvement or 'more easy' to work or more useful…*	

Initial Study 3		Comments
Interviewer:	*Question 14, paragraph A, you selected title 5, so can you tell me why you selected title 5 with reference to the test?*	Attention is drawn to a general statement opening ('background'). The student identifies that the paragraph is about 'why it was established' **[establishing propositional meaning at the clause level]** but didn't perhaps didn't know the word 'prompts' in the correct answer.
Participant:	*Because I think it's just… er… background… about the FAA, that's why I think it's this. The opening sentence… [was the clue]*	
Interviewer:	*How did you know that this was about background information?*	
Participant:	*Because it talk[s] about why it [was] established… um… what this… um… helped to improve air traffic control.*	
Interviewer:	*OK, all right then. What about question 15, paragraph C, you said title 8. OK, Can you tell me why?*	Word matching – 'safety' appears in heading and text **[lexical access]**.
Participant:	*Um… because it mention the safety here [paragraph C, line 6]. It talked about the rules… for… how to reduce the … chance to having accident.*	

Table 3.4. Pilot Studies 2 and 3: verbalisation excerpts and analysis

Verbalisations are now more readily associated with each item in the IELTS test than in initial study 1. However, verbal prompts by the interviewer are limited to asking why the participant selected one specific response, rather than relating to specific moments during the test. Participant verbalisations are still hesitant despite the stimuli. Initial studies 2 and 3 involved participants taking a whole test followed by a single SRI. The first initial study identified a trade-off between authentic test conditions (preserving ecological validity) and designing research to associate verbalisations with specific moments by conducting a series of 'testlets' each with a subsequent SRI. Initial studies two and three demonstrate the utility of a 'testlet' approach for obtaining moment-specific verbalisations. These studies also demonstrate that item-level data may be obtained using the test and responses as stimuli, but moment-specific processing is lost due to a lack of stimuli relating directly to engagement with the test. The initial studies demonstrated that a testlet approach

was beneficial for ensuring data addresses all items individually, rather than item types, but that the proposed stimuli of test, text and responses remained insufficient.

The three initial studies demonstrate that SRI require a principled and robust research design to be in place to address methodological issues associated with SRI. These methodological issues are addressed in the next section. Issues identified in the three studies include the authenticity of the task setting (ecological validity), the nature of the stimuli, the conditions of the SRI, including timing, the language of verbalisations and how participants were selected for the study and the ethical principles that were adopted for this social scientific study.

3.3.3. Methods of data collection of stimulated-recall interviews (SRI)

This section outlines the procedures by which data were collected to address the four research questions. This section begins by combining the issues identified in the initial studies into a single model (figure 3.3) to exemplify how these issues interconnect and affect how claims can be made on the basis of participant verbalisations. Each of these issues is then discussed in more detail.

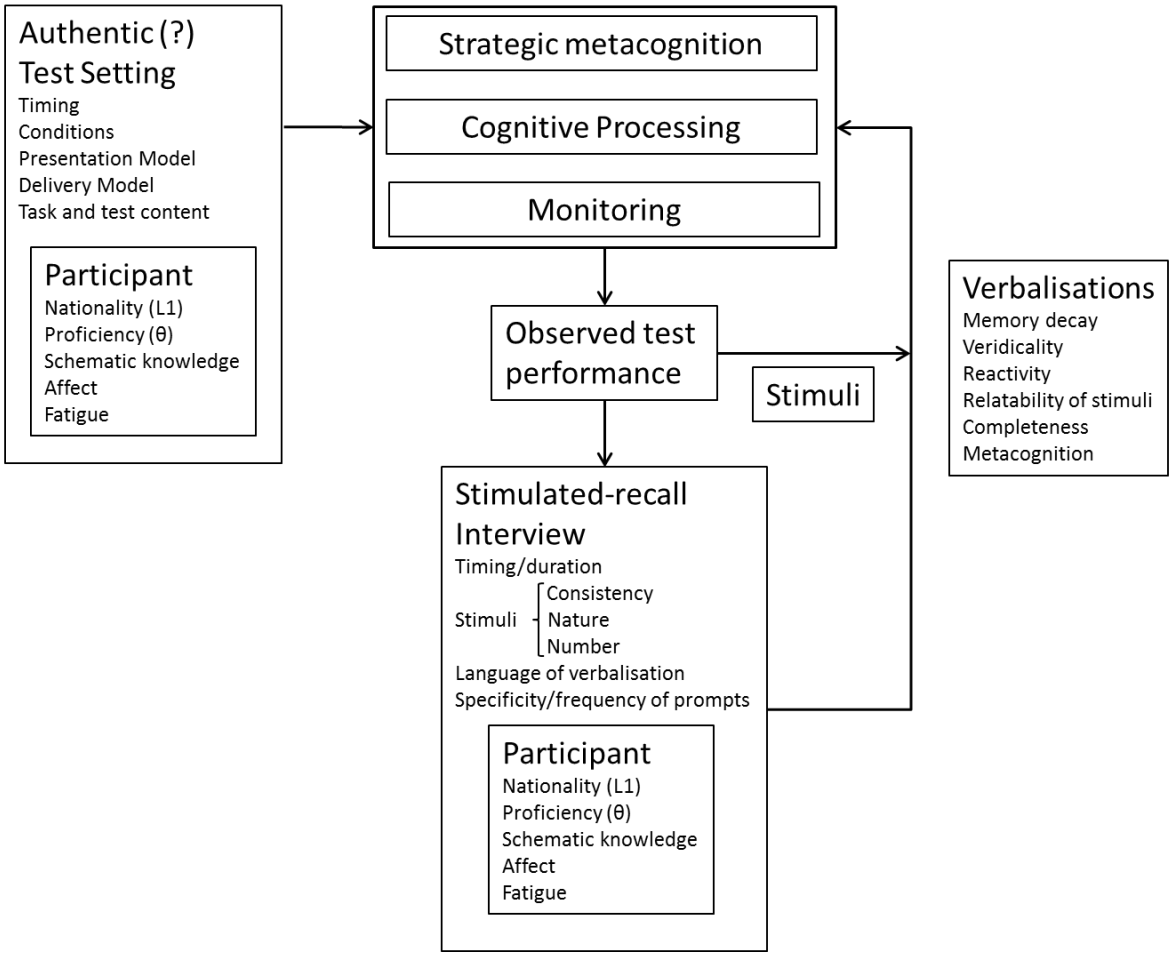


Figure 3.3. Methodological considerations associated with stimulated-recall interviews (Gass and Mackey, 2000; Khalifa and Weir, 2009)

In SRI-based research, a participant is presented with a task in a setting designed to be as authentic as possible. The focus of the participant is solely on the task. The controlled nature of task completion provides evidence for the claim that the participant is providing an authentic performance (behaving as they would in a real high-stakes test). As the participant completes the task, some observational record of the performance is made; either by the participant in the act of completing the task, or some external record by a present researcher or recording device. These records act as stimuli for the participant in the subsequent interview. Once the task is complete, the researcher then initiates an interview with the participant. The conditions of the interview, coupled with characteristics of the participant and the nature of the stimuli will all affect the quality, consistency and duration of the verbalisations and thus the

inferences that can be made from them regarding the cognitive processing the test taker used to select their response. Each of these issues are now expanded upon.

3.3.3.1. Authenticity of the task and setting

In this study, stimulated-recall interviews were preferred to think-aloud verbalisations on the justification of attempting to capture a more authentic task performance. Conditions include a silent environment with minimal distractions. Conditions must be as similar as possible for all participants. Authenticity includes the presentation and delivery models (see Literature Review, sections 2.4.3.1), which complicates a research study with comparability as an agenda. IELTS is paper-based, whilst TOEFL is internet-based, with a specifically designed user interface. As the agenda of this research is comparability, the argumentative case for consistency of conditions across all participants outweighs the ecological validity of maintaining distinct features of the delivery model (the format by which the test is delivered to the participant). The TOEFL test was therefore converted to a paper-based format. However, as many details as possible will remain consistent with the presentation model, including task layout, font size, typeface, item design and spaces between lines. This requires sourcing authentic published test-taking materials and reproducing these as closely as possible for test takers. The selection of tasks and test content cannot be arbitrary. Each test used in SRI research should be a representative example of the developers' test specifications.

3.3.2.2. The nature of the stimuli presented to participants

In the case of a paper-based test of academic reading, readily available stimulus material consists of the test tasks, the text itself and the corresponding answer sheet. This allows a researcher to target individual items specifically, while also maintaining the integrity of the test by requiring participants to complete the test under exam conditions. In the case of video or audio recordings, particular moments can be isolated by the researcher and specifically targeted for verbal feedback if these meet stated research goals.

A further unexplored issue in the literature is the *relatability* of the stimuli. Stimuli must overtly correspond to some element of task performance. If the stimuli are too abstract or too finely-grained and do not exhibit concrete examples of test-taking behaviour, the procedure will negatively impact participants' affective state, specifically their level of motivation. The participant may be uncertain what the researcher wants in terms of verbalisation, resulting in hesitant, descriptive ('I was reading here') or unconfident verbalisations. Mackey et al (2000) created video recordings of participants' interactions in an English speaking test. These were presented to the students, who were also given the remote control and were allowed to control the progression of the recording and to stop at any point they wished to highlight some aspect of their performance. The research agenda of the current study dictates that each item in each reading test needs to be addressed via the observable moments that correspond to it. This way, a stimulated recall procedure will elicit a complete representation of both tests. Therefore, participants were video-recorded as they completed the tests. Their performances were replayed immediately to them upon completion as an additional layer of stimulus.

3.3.2.3. The conditions of the SRI

Conditions of the SRI include the time delay between task completion and the interview, duration of the interview, structure, language of verbalisations and specificity and frequency of prompts and verbal exchanges within the SRI (Mackey et al, 2000). These conditions will affect the willingness of participants to engage in the interview or may induce fatigue. This in turn affects veridicality (Bowles, 2010) and reliability, associated with memory lapses (Gass and Mackey, 2000: 17).

A substantial delay between task completion and the stimulated recall interview will result in memory decay which will compromise the verbalisations, and undermine the validity of subsequent claims made on the basis of the protocols. Similarly, if a task is extremely time-consuming, then the delay between the participant's actions at the beginning of the task and the relevant part of the interview is likely to be substantial.

The initial studies demonstrated the efficacy of a 'testlet' approach; individual mini-tests which may be conducted in a short space of time, followed immediately by a SRI. This procedure may then be repeated two or more times within a single session. Such a procedure would undermine the authenticity of the test setting by removing time constraints. Removing time constraints will have two other implications. First, it may be a positive means of ensuring full coverage of each item. In timed conditions, participants may state that they did not have sufficient time to complete particular items, resulting in lost data. As the participants must be informed that the test is being held in low-stakes conditions, they may not complete the test if timing conditions are too strict. In the present study, full coverage is mandatory to make comparative claims about the two tests and about individual item types. A second implication is that relaxing the time constraints may prompt unrepresentative processing. If participants are aware that they are not being timed, they may decide to read carefully at the global level to maximise their chances of answering all items correctly, rather than reading selectively, as the developers intended. A compromise may be to provide participants with guidelines; inform participants that the test is not timed, but that in an authentic test situation, they should aim to spend approximately ten minutes on these questions.

The level of scaffolding provided for the participant will influence their utterances. A 'high structure' stimulated-recall procedure implies a greater number of highly-targeted interview questions. The converse implies fewer, more general questions. This likely will impact the specificity of utterances and therefore the claims that can be made from this data. As the aim of the research is to build up a cognitive picture of both research instruments at the item and test level, a 'high structure' approach is warranted. However, flexibility in the approach, allowing a mixture of high and low structure verbal prompts, would allow participants to add any information regarding the task or verbal reports that they were unable to share due to the strictures of the previous line of questioning. In terms of individual verbal exchanges, the researchers should not have to model retrieval for the participants to ensure acquisition of relevant information. Verbalisations should be an accurate as possible representation of their thought processes during the task (Ericsson and Simon, 1987: 40-41).

Gass and Mackey (2000: 88) warn that the recall procedure can be disliked or regarded as unnatural by some participants, especially if participants would be verbalising in English which may further impact on participants' affective state. Though participants may stipulate that they are happy to continue and participate, they may refuse to co-operate with the question prompts, yielding unusable data. Instructions must be clear, and participants informed ahead of the test and interview what they are being asked to do and have an opportunity to ask questions to address any uncertainties. Providing diagnostic item-level feedback to participants as part of the study is one effective means of encouraging participation.

3.3.2.4. Language of verbalisations and the selection of research participants

L1 verbalisation was rejected on the basis that it would require translation, which would add a significant time delay to the data analysis and an additional layer of interpretation to the verbalisations. L2 verbalisations determined the proficiency level of potential participants who could be recruited for the study. Lower ability participants were not recruited as they were not deemed to be able to verbalise their item completion processes. An inability to verbalise thought processes will result in incomplete verbalisations, meaning that no statements regarding participants' strategy use or cognitive processing could be made. Objective evidence of participants' proficiency in the form of test scores (for both reading and speaking) were collected as evidence that participants were able to take part in the research.

Six postgraduate Chinese students from the University of Leicester were recruited for the study. They were informed of the nature of the study but not the specific research questions, coding schema or the aims to identify specific strategic and cognitive processes. Each participant was informed that they would need to participate in two sessions. All twelve interviews occurred between July – August 2014. As none of the participants had taken the TOEFL test before, they were sent a practice test via email one week before they were due to participate in the TOEFL test and interview. They were instructed to familiarise themselves with the item types prior to arriving for the

real test. Additionally, before the test began, all item types were shown to the participant and verbally explained. Participants then had time to ask questions if they were uncertain about any elements of the test.

Participant	L1	Test taken	Overall score	Reading Score	Speaking Score	Date of administration	Academic Pathway
1	Chinese	IELTS	7.5	7	7	03/2013	MA TESOL
2	Chinese	IELTS	7	7.5	7	06/2012	PhD Education Research
3	Chinese	IELTS	8	8.5	7.5	02/2007	PhD Education Research
4	Chinese	IELTS	6.5	6.5	7	03/2012	MSc Financial Economics
5	Chinese	IELTS	7	7.5	6	06/2013	MA Media and Public relations
6	Chinese	IELTS	6.5	6	6	08/2011	MSc Financial Economics

Table 3.5. Evidence of linguistic proficiency of research participants

All six participants were Chinese. The research design prioritised depth of cognitive processing rather than breadth. Therefore, the number of participants required in the study could not account for all of the linguistic variation that represents the test taker population. Therefore, the study will not attempt to make generalizable claims regarding the test taking population. Claims will be limited to those that can be made about strategies and processes used by the participants for the selected test instruments. L1 background was seen as an intervening variable to be controlled in the study. Therefore, the research commenced with six participants from the same linguistic background. The number of Chinese students studying at UK universities is increasing year on year (Higher Education Statistics Agency, 2014),

	2009/10	2010/11	2011/12	2012/13	2013/14
Total	56,990	67,325	78,715	83,790	87,895

Table 3.6. Number of Chinese students at all UK HEPs (https://www.hesa.ac.uk/free-statistics)

The number of overseas students at UK HEPs is increasing and in 2014 Chinese students accounted for 28% of the total (ibid.). It was therefore decided that Chinese participants would reflect this contemporary trend. Each of the participants had gained the requisite score to undertake a postgraduate programme at the University of Leicester and were therefore deemed suitable participants. As the study involved recruitment of students, there were several ethical procedures to follow, which are outlined in the next section.

3.3.2.5. Ethical procedures and implications

The ethical approaches to the study were considered using Stutchbury and Fox's (2009) ethical appraisal framework. The authors recommend constant appraisal throughout a research project based on four elements; focus (consequential), approach (ecological), methods (relational) and obligations (deontological). Each of the four dimensions was appraised when planning, conducting and reporting the study.

The study received ethical approval from the University Ethics panel in May 2013. The study required the consent of six participants, each of whom received a consent form and instructions detailing the level of involvement and the stimulated-recall procedure (see Appendix A). The procedure was verbally explained to participants again in the interview. Participants were informed that they were able to withdraw at any time and that if they wished to withdraw, their information would not be retained.

Piloting revealed a potential conflict between relational and deontological thinking. Initially, the research was advertised on the premise of a test which would last approximately 20 minutes, followed by an interview of approximately 20 minutes. However, as the test was divided into testlets, individual interviews lasting longer than predicted and time required for transferring video to a computer for viewing in the interview, each session lasted between 1.5 and 2 hours. Consequently, the information and consent forms were amended to better reflect the increased level of commitment (deontological research obligation), at the risk of putting off potential participants, requiring a renewed emphasis on relational ethical thinking to persuade the

participants to give up so much of their valuable time. A more face-to-face and personal recruitment strategy was then planned instead of advertising to the sider student community via email and posters.

The study requires the utilisation of copyrighted material from two testing companies (Cambridge Assessment and Educational Testing Service). Cambridge Assessment gives permission to publish short prose extracts from their publications of less than 400 words[1]. As the research project will not be published in its current form, permission was not needed to use these materials to and quote from them in the findings and discussion chapter. Fees for materials used in research are waived for research purposes by ETS[2]. The participants were not permitted to retain any data or information related to the tests after they had completed the stimulated-recall task. All test-related information was retained by the researcher.

On 18th April 2014, the researcher was notified that the research was to be formally supported in the form of an ETS Small Doctoral Award of US$2000. From this fund, participants were each paid £15 in cash as remuneration for their participation in the research. This was paid to them after they had completed the stimulated-recall interview.

3.3.2.6. Finalised data collection procedure

Two tests (one IELTS and one TOEFL iBT) were selected. The test selection procedure is outlined in section 3.3.2.6. As the initial studies and the pilot study demonstrated the efficacy of a 'testlet' approach, each of the tests were divided into three parts, consisting each of a single text and associated items. Each testlet was administered in an independent session. Each test required two complete SRI records for comparative purposes. Therefore, twelve independent sessions with six participants were conducted, as represented in table 3.7:

[1] http://www.cambridge.org/about-us/rights-permissions/permissions/permissions-requests/
[2] http://www.ets.org/legal/permissions/licensing#toefl

	Session 1 (Test 1)	Session 2 (Test 2)
Participant 1	IELTS 1 (Q1-13)	TOEFL 1 (Q1-13)
Participant 2	IELTS 2 (Q14-26)	TOEFL 2 (Q14-26)
Participant 3	IELTS 3 (Q27-40)	TOEFL 3 (Q27-38)
Participant 4	TOEFL 1 (Q1-13)	IELTS 1 (Q1-13)
Participant 5	TOEFL 2 (Q14-26)	IELTS 2 (Q14-26)
Participant 6	TOEFL 3 (Q27-38)	IELTS 3 (Q27-40)

Table 3.7. Research design for participants and tests

Participants 1-3 completed the IELTS testlet in their first session, with the TOEFL testlet in their second session. Participants 4-6 complete the same testlets, but in the reverse order to account for any method effect, such as participants becoming more used to the procedure and providing more in-depth data in their second sessions (participants 'learning' how to verbalise). Each session was further divided into segments of 3-5 items each. Participants were asked to complete a small number of items, then participate in a SRI. Neither the tests nor the interviews are timed, providing as much flexibility for the researcher and the participants as possible.

As they completed the test, participants were video-recorded, with the camera aimed at their hands and test papers. When the participant was satisfied they had finished each segment, they informed the researcher, who was seated next to the participant. The camera was connected to a laptop computer and the video of the performance immediately uploaded. Then the SRI commenced. Participants verbalised in English. Evidence of English language proficiency in the form of speaking scores was added to the information to be gathered for to provide evidence that participants are capable of explaining their thought processes in depth.

The video, participant responses, annotated questions and text formed the stimuli. Verbalisations were stimulated by specific actions displayed on the screen, annotation or item responses. If an action is unclear on the video, it is possible to look at the text or question paper to discern what action had been taken (e.g. underlining specific words, or writing a specific response). Progress through the video was determined by the participant. They were able to pause, rewind and replay any specific moments of their choosing. The interview terminated when the video recording finished.

Verbalisations were captured using a digital audio recorder. Verbalisations that included deictic language ('this', 'that') was accompanied by a short explanation by the researcher indicating the object or location of the reference. Recordings will be transferred to a personal computer, allowing easy navigation between portions of the recording using Windows Media Player™. Transcriptions occurred as soon after the interviews as possible to ensure an accurate representation of participants' verbalisations.

Tests and interview sessions were conducted between August and September 2014. Twelve test and interview sessions were held with six participants. Video and interview recordings varied considerably depending on participant actions and the amount of explanation they were able to provide in interview. Participants had freedom to determine where they split their tests into testlets. Participant 1 for example, chose to have four individual testlets in their first session, broken up according to item type in part 1 of the IELTS test. Video recordings were made on a Panasonic Lumix TZ7 Digital Camera, recording high definition at 720p. An Apple iPad Mini was used as back-up camera in the event of battery or technical failure. For four of the videos it was necessary to use the iPad and this revealed two advantages. Video replay on an iPad is instantaneous and can be displayed full-size. Video recorded on a digital camera needed to be uploaded to a laptop computer, taking approximately 2 minutes per video. Video recording on the iPad was also higher quality – full 1080p high definition. These videos were much clearer in terms of identifying participant actions in relation to specific parts of the text. In some instances, it was necessary to clarify actions taken by the participant during the interview. Drawbacks with using the iPad were the limited memory capacity of 16GB, (meaning each video could not exceed more than 20 minutes) and the difficulty of uploading video from the iPad hard drive to a laptop computer. Nonetheless, for the purposes of future analysis, I recommend the use of iPads for the beneficial feature of instant video playback as this is crucial to counter the argument of incomplete or inaccurate verbalisations due to memory decay, and a clearer picture to more easily identify participant actions.

Participant	Session 1 (Test 1)	Date	Video Duration	Interview Duration	Session 2 (Test 2)	Date	Video Duration	Interview Duration
1	IELTS 1 (Q1-13)	01/08/14	00:09:04 00:04:42 00:06:26 00:08:08	00:36:22 00:06:29 00:02:14	TOEFL 1 (Q1-13)	25/08/14	00:12:10 00:20:14	00:28:30 00:30:46
2	IELTS 2 (Q14-26)	04/08/14	00:19:06 00:10:50	00:40:26 00:24:42	TOEFL 2 (Q14-26)	22/08/14	00:03:35 00:04:49 00:20:03	00:14:05 00:15:33 00:44:08
3	IELTS 3 (Q27-40)	23/08/14	00:09:41 00:04:57	00:25:19 00:22:38	TOEFL 3 (Q27-38)	03/09/14	00:12:28 00:06:24	00:24:40 00:19:41
4	TOEFL 1 (Q1-13)	25/08/14	00:18:04 00:03:27 00:17:21	00:26:26 00:31:56	IELTS 1 (Q1-13)	28/08/14	00:12:01 00:12:58	00:14:13 00:15:55
5	TOEFL 2 (Q14-26)	27/08/14	00:11:06 00:13:44	00:23:12 00:21:08	IELTS 2 (Q14-26)	02/09/14	00:07:57 00:13:52	00:14:44 00:02:04 00:22:14
6	TOEFL 3 (Q27-38)	27/08/14	00:12:31 00:08:32	00:19:41 00:18:29	IELTS 3 (Q27-40)	01/09/14	00:10:43 00:07:03	00:14:12 00:11:44
Total time (hrs/mins/secs)			**2:37:39**	**4:59:02**	**Total time (hrs/mins/secs)**		**2:24:17**	**4:32:29**

Table 3.8. Test-taking and interview recording times

Total video recording time for the tests exceeded five hours (5:01:56). Subsequent interviews for the six participants exceeded nine and a half hours (9:31:31). Interview comments were recorded using a digital voice recorder. Once the videos and recordings were uploaded, participant verbalisations were transcribed in Microsoft Word™ and aligned to specific moments in the video recordings. Interviewer and participant comments were transcribed in full. Transcription occurred throughout September 2014. Due to the volume of data collected, transcriptions were made up to one month after the interviews had taken place. This problem may have been alleviated by the recruitment of transcribers. Nonetheless, transcription was undertaken by the researcher for two reasons. First, the data is highly contextualised.

Specific moments in the recordings may refer to specific items or portions of a text. This may cause problems for transcribers who are not involved in the research. Secondly, transcription was viewed as a means of immersing myself in the data. Issues that arose during transcription could be noted for analysis later. The videos were viewed again and descriptions written for observable actions taken by the participants, including when they answered items and which response they selected. This information was entered into the following data matrix:

Observable action (video)	**Action Code(s)**	**Highest level of processing inferred**	**Verbal cue (transcript) 01:01:01:1-13**
TIME Action 00:00 Description of action from video	Strategic action	Cognitive processing inferred	Participant verbalisations

Table 3.9. Analysis framework for participant verbalisations

Moments in the left-hand column refer to observable actions and the specific minute and second in the video recording that they occurred. In the second column, these actions were coded either on the basis of observable actions alone, or supplemented with an understanding from the verbalisations in the right-hand column. Verbalisations were aligned to specific observable moments in the video recordings. The remaining column would be used to code individual verbalisations with the highest level of processing that could be inferred (using Khalifa and Weir's (2009) model) based on participants' statements made in reference to observable individual moments in the test.

The IELTS test contains forty items. The test was divided into three parts, each related to a single text. Each of the three parts was completed twice by individual participants for comparative purposes. Verbal records were obtained for eighty items (two verbal records for each item).

3.3.2.7. Selection of the IELTS and TOEFL tests for data collection

The strength of the claims regarding the cognitive processing in IELTS and TOEFL rests upon the assumption that the instruments selected for the research are representative examples of live tests. It therefore became necessary to develop an objective approach to instrument selection, which this section details. The chosen IELTS and TOEFL tests represent complete and authentic reading tests. Coh-Metrix[3] and VocabProfile (Cobb, 2013)[4] were adopted to assist with the task of instrument selection. These are online, free to use, text analytic software. They produce a range of measures of readability, ranging from simple descriptive metrics (number of words, number of paragraphs, type-token ratios) to more sophisticated measures of cohesion and coherence using latent semantic analysis (LSA). LSA is a global co-occurrence model (Vigliocco and Vinson, 2007) that defines conceptual similarity between words on the basis of their closeness in any linguistic context. It therefore assumes that words which co-occur in any language are more closely related semantically than those which do not frequently co-occur.

The tests used in the research would be those which contain texts which best represent both IELTS and TOEFL. To do this, 24 texts (12 TOEFL and 12 IELTS) texts were scanned directly from authentic example tests. These texts represented eight complete reading tests, four TOEFL and four IELTS, each from different, official test preparation materials produced by Educational Testing Service (ETS) and University of Cambridge ESOL examinations respectively. These scanned documents were entered into text recognition software (Microsoft OneNote 2010™). They were then read carefully to check for errors against the source materials. Spelling, punctuation, grammar and paragraph splits remained consistent with original sources, regardless of whether the sources used US or UK spelling conventions. These texts were then entered into Coh-Metrix v3.0 and VocabProfile to produce data for each text.

[3] http://www.cohmetrix.com/
[4] http://www.lextutor.ca/vp/comp/

Coh-Metrix data was entered into SPSS v.22. Each text represented a single case. The tests were assigned memberships to one of two groups (0 = TOEFL, 1 = IELTS). Each of the 108 metrics from Coh-Metrix and nine metrics from VocabProfile represented independent scale variables, with 'test' as a nominal variable. A series of two-tailed t-tests (recognising that either IELTS or TOEFL could record higher measures for each variable) were performed on the data to identify differences between the groups. To mitigate the propensity for type I error, statistical significance was interpreted at p<.01. Fourteen of the 117 metrics met this threshold:

Measure	F	Sig.	t	df	Sig. (2-tailed)	Cohen's d	99% CI	Power (%)*
Paragraph count, number of paragraphs	0.72	0.40	-3.19	46.00	<0.001	-0.92	0.24	73.30
Sentence count, number of sentences	6.25	0.02	-4.44	38.02	<0.001	-1.28	0.26	96.90
Word count, number of words	9.48	0.00	-8.15	27.70	<0.001	-2.35	0.36	100
Text Easability PC Referential cohesion, z score	2.07	0.16	2.73	46.00	0.01	0.78	0.23	55.30
Text Easability PC Verb cohesion, percentile	1.37	0.25	-2.60	46.00	0.01	-0.75	0.23	50.90
Stem overlap, all sentences, binary, mean	0.77	0.38	2.74	46.00	0.01	0.75	0.23	51.50
LSA overlap, adjacent sentences, mean	0.13	0.72	4.37	46.00	<0.001	1.27	0.26	96.50
LSA overlap, adjacent sentences, standard deviation	0.78	0.38	2.57	46.00	0.01	0.67	0.23	39.50
LSA overlap, all sentences in paragraph, mean	0.04	0.85	4.53	46.00	<0.001	1.40	0.27	98.90
LSA overlap, adjacent paragraphs, mean	0.76	0.39	4.58	46.00	<0.001	1.41	0.27	98.90
LSA given/new, sentences, mean	0.23	0.63	4.77	46.00	<0.001	1.33	0.26	97.90
Noun incidence	1.98	0.17	2.91	46.00	0.01	0.84	0.23	62.90
First person plural pronoun incidence	10.32	0.00	-2.85	26.55	0.01	-0.82	0.23	60.60
Familiarity for content words, mean	0.01	0.94	-2.59	46.00	0.01	-0.75	0.23	50.80

Post-hoc calculation

Table 3.10. Effect size and statistical power for fourteen Coh-Metrix measures which showed statistical differences between IELTS and TOEFL.

Means and standard deviations were also calculated for each of these metrics for both IELTS and TOEFL. The tests chosen would be those which contained texts for which the majority of the metrics were located *within* one standard deviation of the mean value of that metric. Thus, that test would then be considered a 'representative' example of that test. Metrics which produced statistically significant differences between IELTS and TOEFL were chosen in order to ensure that any differences in the test development process were preserved in the instruments. The extent to which the metrics included in table 3.7 affect strategic actions or cognitive processing is unexplored. Nonetheless, preserving any changes to the texts made in the course of test development ensures that the tests can be claimed to be authentic representations of both IELTS and TOEFL. Of the eight IELTS tests analysed, test 5 was selected. Of the 42 measures (14 per text), only six fell outside the threshold of one standard deviation away from the mean for that metric. Of the eight TOEFL texts, the second was chosen and recorded eight measures outside of one standard deviation from the mean for those metrics.

A summary of the selected tests is outlined in tables 3.11 and 3.12 below. The tables contain the title of each text, the number of items and item types associated with each text, and the abbreviations which will be used to identify individual item types:

IELTS		
Text	**Items**	**Question type and claimed strategies/processes**
"Striking back at lightning with lasers"	1-3	**Type 1: Multiple-choice (1-MC)** Detailed understanding of specific points; general understanding of the main points of the text.
	4-6	**Type 8: Sentence completion (8-SC)** Ability to find detail/specific information in a text.
	7-10	**Type 9: Summary completion (9-SuC)** Ability to understand details and/or the main ideas of a part of the text; type of word(s) that will fit into a gap (for example, whether a noun is needed, or a verb, etc.).
	11-13	**Type 2: Identifying information (Yes/No/Not given) (2-IDi)** Ability to recognise specific information given in the text.

"The Nature of Genius"	14-18 19-26	**Type 4: Matching information (4-MI)** Ability to scan a text to find specific information. **Type 2: Identifying information (True/False/Not given) (2-ID)** Ability to recognise specific information given in the text.
"How does the biological clock tick?"	27-32 33-36 37-40	**Type 5: Matching headings (5-MH)** Ability to identify the general topic of a paragraph; recognise the difference between the main idea and a supporting idea. **Type 8: Sentence completion (8-SC)** Ability to find detail/specific information in a text. **Type 3: Identifying writer's views/claims (3-IDw)** Ability to recognise opinions or ideas.

Table 3.11. Item types (and codes) for IELTS instrument
http://www.ielts.org/test_takers_information/question_types/question_types_-_ac_reading.aspx

The IELTS test contains three texts and 40 questions. It is noteworthy that the IELTS test selected for the research does not contain the full range of item types associated with the IELTS reading test (item types 6 and 7 are *not* represented). Nonetheless, this is common practice in IELTS tests. Not all reading tests contain the full range of item types that Cambridge has developed. As a full-length test was utilised, the argument that it is representative of a live IELTS test can be made. Decisions are made about test takers in high-stakes testing situations without them responding all possible item types. Therefore, an implicit claim is made by Cambridge ESOL that reliable and valid claims can be made about test takers on the basis of a *selection* of item types. Therefore, there must be overlap in terms of the construct representation by each of the item types, and that the above instrument can still represent a full range of construct-relevant cognitive processing. This will be investigated as part of research question 4. The selected TOEFL test is outlined below in table 3.12:

TOEFL			
Text	**Items**	**Question type and claimed strategies/processes**	
"19th Century Politics in the United States" (Q1-13)	1; 6; 8	Vocabulary question (BC-v)	**Reading to find information** Recognise and identify words efficiently. **Basic comprehension** Comprehend the essential meaning of major propositions; efficiently comprehend major propositions.
	2; 4-5; 7	Factual information (BC-f)	
	10	Negative factual information (BC-nf)	
	11	Sentence simplification question (BC-ss)	

	3	Rhetorical purpose question (I-rp)	**Inferencing** Integrate and remember major ideas and supporting information; organise important information; express major ideas and supporting ideas.
	9	Inference question (I)	
	12	Insert text question (I-it)	
	13	Prose summary (R2L-ps)	**Reading to learn** Generate organisational framework and relate ideas and information from two or more sources.
"The Expression of Emotions" (Q14-26)	14; 16; 22-23	Vocabulary question (BC-v)	**Reading to find information** Recognise and identify words efficiently **Basic comprehension** Comprehend the essential meaning of major propositions; efficiently comprehend major propositions.
	17	Pronoun reference Question (BC-pr)	
	18; 20-21; 24	Factual information (BC-f)	
	19	Sentence simplification question (BC-ss)	
	15	Rhetorical purpose question (I-rp)	**Inferencing** Integrate and remember major ideas and supporting information; organise important information; express major ideas and supporting ideas.
	25	Insert text question (I-it)	
	26	Prose summary (R2L-ps)	**Reading to learn** Generate organisational framework and relate ideas and information from two or more sources.
"Geology and Landscape" (Q27-38)	28; 30; 33	Vocabulary question (BC-v)	**Reading to find information** Recognise and identify words efficiently. **Basic comprehension** Comprehend the essential meaning of major propositions; efficiently comprehend major propositions.
	27; 31; 36	Factual information (BC-f)	
	34	Pronoun reference question (BC-pr)	
	35	Sentence simplification question (BC-ss)	
	29	Inference question (I)	**Inferencing** Integrate and remember major ideas and supporting information; organise important information; express major ideas and supporting ideas.
	32	Rhetorical purpose question (I-rp)	
	37	Insert text question (I-it)	
	38	Schematic table (R2L-st)	**Reading to learn** Generate organisational framework and relate ideas and information from two or more sources.

Table 3.12. Item types (and codes) for TOEFL instrument (Cohen & Upton, 2006; Jamieson et al., 2008)

The selected TOEFL test contains three texts and 38 questions. Each part contains examples of all of the item types that ETS designed, with comparable numbers of each item type in each part of the test. Each part contains basic comprehension, inferencing and reading to learn items. The majority in each are basic comprehension items (8-10). Each test contains one reading to learn item, either a schematic table or prose summary. Contrasted with IELTS, this suggests that ETS has specific ideas about which cognitive processes are associated with specific item types. As with IELTS, claims made about each item type (Cohen and Upton, 2006; Jamieson et al., 2008) are used in the findings and discussion chapter as a means of comparing claims made by the developers against emergent data.

This section has outlined the procedure by which data will be collected in the SRI. The next section explicitly outlines how the collected data was analysed, focusing on how Khalifa and Weir's model of reading (2009) was transformed into a coding scheme, and how this coding scheme was reliably applied to the verbal data. Also included in this section is the details of a pilot study that was conducted to ensure that appropriate data could be collected with the finalised procedure and could be analysed using the coding schema.

3.3.4. Methods of data analysis

The methodology prioritises depth of engagement with individual tests rather than surface-level engagements across multiple test versions, as the research questions require highly-detailed data to be able to make claims relating to test and individual items. This section will also present the coding schema and how the comparative approach to the IELTS and TOEFL tests would be undertaken using the schema developed from Khalifa and Weir's (2009) reading framework.

3.3.4.1. Establishing a coding schema from Khalifa and Weir's (2009) model of reading

Visual data from the video recordings (strategic actions) and the transcribed verbal data (cognitive processes) were both coded. Identification of strategic actions occurred organically from the video. All participant actions that could be observed in the video were noted, defined and added to the rubric. A strategy rubric was built from observing participant behaviour across the video stimuli. As more videos were watched, the number of strategies in the rubric increased. A tally of each strategy was kept as the videos were watched. Strategies were identified as means of engaging with the input material, with the recognition that participants may then engage with the material at a deeper level than is observed in the recording. There was no fixed number of strategies to be identified – the number of strategies in the final rubric would be determined by the number that could be identified in the video recordings. The strategy rubric is presented as part of the findings and discussion for research question 1.

The rubric for cognitive processing is formed from the reading model presented by Khalifa and Weir (2009). Individual cognitive processes were identified from participant verbalisations given during the SRI. The verbalisations provide the basis for determining the level of cognitive processing that the participant must have performed to be able to verbalise that understanding. Table 3.13 below demonstrates how Khalifa and Weir's (2009) model was adapted to form a coding schema for participant verbalisations:

	Level of processing	**Code**
Higher-level processes	**Creating an intertextual representation**: construct an organised representation across texts	**[P8]**
	Creating a text-level representation: construct an organised representation of a single text	**[P7]**
	Building a mental model: Integrating new information; enriching the proposition	**[P6]**
	Inferencing: At word/sentence/clause level	**[P5s/c]**
	Inferencing: At word level	**[P5w]**
Lower-level processes	**Establishing propositional meaning**: At sentence level	**[P4s]**
	Establishing propositional meaning: At clause level	**[P4c]**

	Syntactic parsing	**[P3]**
	Lexical Access	**[P2]**

Table 3.13. Final coding scheme for processing core of Khalifa & Weir's (2009) model of reading

The lowest category (word recognition [P1]) was removed from the model. Coding each instance of word recognition would obscure other, higher-level processes and would not provide adequate construct information to be useful to stakeholders, and was not considered theoretically interesting. Word recognition encompasses orthographic recognition – that is, recognition of specific letters which form particular words. It was not considered likely that a participant would verbalise that they selected a response on the basis of recognising a word in a question stem and identifying the same word in the text without having any understanding of that lexical item, which would instead be coded as 'lexical access'.

Lexical access (P2) refers to instances in which participants select a response on the basis of identifying key words in item stems, options and finding identical or equivalent lexis in the text. Syntactic parsing (P3) refers to verbalisations in which participants state that function words assisted them in getting the correct response; the sentence structure revealed something about the intention of the item that led them to select a response. 'Building a mental model' (P6) refers to verbalisations that summarise portions of the text and demonstrate that the participant was able to differentiate main form supporting details. 'Creating a text-level representation' (P7) is more difficult to code. It requires participants to verbalise understanding of several main points in the text and the purpose of the text itself. Text-level representation is unlikely to reveal itself in single verbalisations. This code was used when participants have verbalised several instances of building a mental model and then demonstrated text-level understanding when they were able to clearly state what the purpose of the text is and can locate main arguments throughout the text. It was not anticipated that this code would be utilised much; it was used to test claims made by test developers that specific item types require participants to form a text-level representation, or whether participants were able to respond correctly using lower levels of processing.

Inter-textual representation (P8) requires evidence of forming text-level representation across two texts, and then elaborating on how their understanding of one has informed the other. As each text in IELTS and TOEFL are designed to be discrete, knowledge of one text influencing how participants respond to an item relating to another text is regarded as construct-irrelevant variance. Nonetheless, this is an important component of the construct. An absence of this skill implies that either or both tests have a limited conception of reading in English for academic purposes.

Two significant changes were made to the model for coding purposes. 'Establishing propositional meaning' (P4) and 'inferencing' (P5) were divided into two codes each. Level four ('establishing propositional meaning') in the model is cited by the authors as occurring at the *clause* and *sentence* levels (2009: 43). It became clear from participant verbalisations in the initial studies that instances in which a participant responded on the basis of understanding clause-level information could be distinguished from responses made from establishing propositional meaning of multi-clause sentences. The latter are regarded as more complex than the former as the participant is required to process more information.

Despite 'inferencing' being a higher-level process in the model, this process only requires the linking of two *lexical items* in adjacent (or non-adjacent) sentences. However, the further the distance between the noun and its pronoun, the more complex the cognitive processing is likely to be. Additionally, more complex inferencing involves mentally linking propositional ideas across sentences which are not explicitly linked in the text. This suggests that inferencing should be split into two internal categories; 'inferencing at the word level' and 'inferencing at the sentence level'. Coding verbalisations accordingly adds an additional dimension to test comparison to determine the *nature* of inferencing in the two tests.

Inferencing (P5) is therefore separated into two categories (inferencing at the word [P5w] and clause/sentence level [P5s/c] respectively). Inferencing at the word level refers to instances of participants citing a need to integrate meaning across sentences at the word level, such as linking a pronoun to its referent (anaphoric and cataphoric

referencing). Cataphoric referencing (referring to pronouns in previous sentences) is cited by Khalifa and Weir (2009: 51) as evidence of inferencing as it involves mentally linking information which is not explicitly stated across two sentences. Anaphoric referencing (linking a pronoun to its constituent in a *subsequent* sentence) may also be considered a form of inferencing despite not being mentioned by the authors. Instances of this were also coded 'P5w'. Inferencing at the clause or sentence level refers to instances in which participants establish meaning in two or more adjacent or non-adjacent sentences (each of which may contain one or more clauses) and then verbalising a proposition which is formed from understanding of those two clauses or sentences which is not explicitly stated in the text. Separating these codes was regarded as a means of conducting more finely-grained comparative analysis of tests and item types.

Coding was conducted by the researcher as the research project involved the development of two unique coding schemes for both strategic and cognitive processing. Subtle distinctions between the levels deserve close scrutiny in order to justify comparability of findings, such as relating to levels of inferential reasoning outlined above, which would not be possible with the recruitment of multiple coders. As the model has not previously been adapted for use as a coding scheme, part of the discussion focuses on the success of the model to identify instances of lower and higher-level processing, and the completeness of the coding scheme in terms of how it is able to account for all of the verbalisations made by the participants. Nonetheless, the lack of multiple raters means that an alternative argument is required to demonstrate that the coding rubric was applied consistently to the verbalisations. For this reason, a coding algorithm was created to aid rule-based decision-making about individual verbalisations. This is elaborated in the next section.

3.3.4.2. Application of the coding scheme and the development of a coding algorithm to ensure reliability of coding participant verbalisations

Each verbalisation was coded with the highest level of processing that could be inferred by the researcher. Khalifa and Weir's model is already 'algorithmic' in nature,

with success at one level of processing presented as a necessary condition for the next level of processing to occur. Coding each verbalisation with all relevant levels of processing would skew the findings towards a preponderance of lower-level processes. Additional coders were not recruited for the study. Research question 2 partly aims at discerning how Khalifa and Weir's model could be used to identify the relevant cognitive processes used by test takers when engaging with text and different item types. Therefore, the applicability of the model and how explicit verbalisations by participants were coded consistently formed part of the findings for this research question. To aid this process, an algorithm was developed in a grounded approach of engaging with participant verbalisations, comparing them to the video evidence and participant responses to test items.

The development of this algorithm facilitated the transition of Khalifa and Weir's processing core from a model of reading to a research instrument. The algorithm is presented in figure 3.4 below. The process of coding begins in the top left-hand corner. Observable participant behaviour guides the initial 'yes/no' decision. If participants are observed engaging at the word level, they are prompted to explain their thoughts. If they cite word-level information, this verbalisation is coded P2. If they provide evidence of syntactic parsing, the verbalisation is coded P3. If their verbalisation indicates that they used the identified lexical items to establish propositional meaning associated with those key words, then it is coded P4c. Note that at this point, the algorithm merges with a negative response to the initial question. If there is no evidence of word-level engagement, such as prolonged careful reading without marking any part of the text, the participant is asked for their thoughts at that moment. If they are not able to cast their mind to that moment, then no coding is assigned. If during careful reading, the participant is able to offer prolonged explanation of what they were reading and why they were reading it, the coder is then tasked with categorising this verbalisation according to the level of detail offered by the test taker (P4c – P8). Explicit elaboration of how the coding scheme and algorithm was applied to verbalisations forms part of the findings. This algorithm was applied to each verbalisation that was gathered in the data collection to form the basis of the

cognitive specification matrices for both the IELTS and TOEFL items. The coding algorithm is reproduced below in figure 3.4:

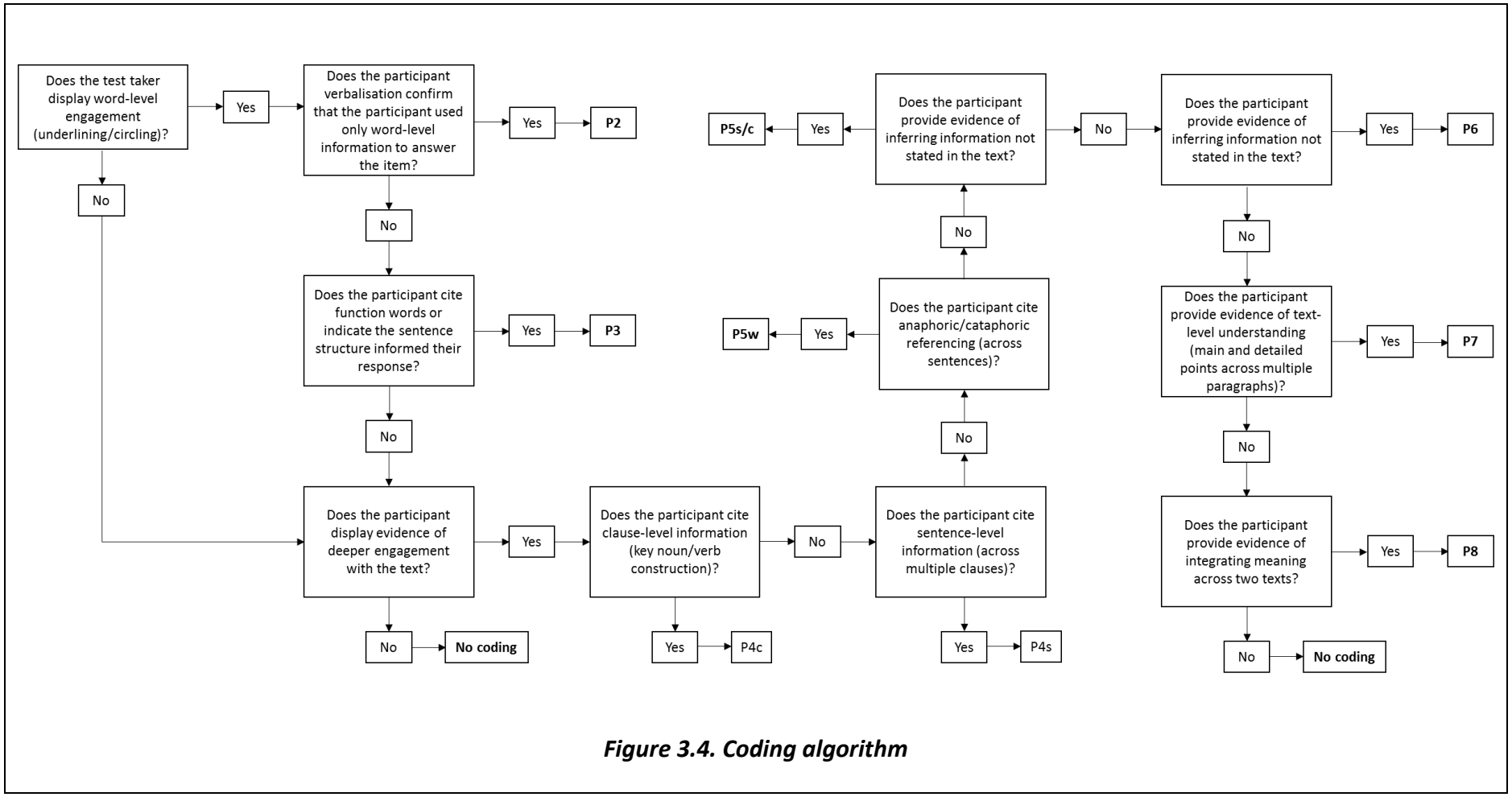


Figure 3.4. Coding algorithm

3.3.4.3. How coded data will be used to compare cognitive processing in IELTS and TOEFL

The research questions mandated the collection of highly-detailed data for comparative purposes across tests and item types within the tests. A coding strategy that is highly amenable to quantification was therefore adopted for this purpose. However, a method of comparing the outcomes beyond simple addition was required. Cohen and Upton (2006) devised an extensive, three-part rubric for reading and test-taking strategies which was replicated for individual item types for both IELTS and TOEFL in this study. This was done to provide comparability of findings in this study with those of Cohen and Upton and also to provide the comparability measure across item types and tests. Cohen and Upton noticed that due to the different frequency of item types in the instruments, a simple occurrence measure would not accurately report the relative importance of each strategy, as some strategies were reported multiple times in relation to a single item type. The authors developed a simple ratio, whereby the number of occurrences of each strategy was reported in relation to the number of that item type. This provided an impression of the relative importance of that strategy (Cohen and Upton, 2006: 41):

Very high (VH) frequency	≥ 1.00
High (H) frequency	≥ 0.50
Moderate (M) frequency	≥ 0.30
Low (L) frequency	≤ 0.29

This scale was adopted in the current study for reporting of strategies that emerged in response to the test performance in both the TOEFL and the IELTS and the cognitive processes that were inferred from participant verbalisations. However, the number of strategies that were observed necessitated a more finely-grained approach. The ratio was adapted in the current study to provide greater nuance. The metric was modified by the addition of another level. 'Very high frequency' (VH) was used for instances of a strategy occurring at a metric value of greater than or equal to two. Between one and two became 'high frequency' (H). 0.5 – 1 became 'moderate frequency' (M), 0.3 – 0.5 became low (L) and

0.29 and below became 'sporadic' (S), indicating that this strategy was used infrequently with no discernible regularity in its usage:

Very high (VH) frequency	≥ 2.00
High (H) frequency	≥ 1.00
Moderate (M) frequency	≥ 0.50
Low (L) frequency	≥ 0.30
Sporadic (S) frequency	≤ 0.29

Table 3.14. Final scale weighting of frequency of strategies divided by the number of items per item type (adapted from Cohen and Upton, 2006)

This data was then presented in the form of a cognitive specification matrix, with items on the x-axis and cognitive processes on the y-axis, identifying which processes are most important in relation to each item type across both IELTS and TOEFL.

The methods of data collection and analysis were put into practice with a single pilot study, which was carried out in order to ensure that the adopted methodology was able to capture the relevant verbal data. Details and outcomes of the pilot study are outlined in the next section.

3.3.4.4. Pilot study

The pilot study was conducted in line with the data collection procedures outlined in section 3.3.3. The pilot study was conducted using a TOEFL instrument to determine whether a similar approach yielded usable data with the TOEFL test, as the three initial studies had demonstrated that the IELTS test was amenable to the data collection procedure. The participant was a male doctoral student studying Management who had successfully passed their viva voce examination with minor corrections prior to participating in the research. The participant had not previously taken the TOEFL test. For this study, the video stimulus for participant and researcher was employed. The participant was filmed as he completed the test, with the camera focused on their actions. A video still of the recording is included below (see figure 3.5). The video was replayed to the participant, allowing him to focus verbalisations on specific moments in the test. This stimulus was also used to prompt more

detailed reports. The participant's face was not filmed. The camera was aimed at their test paper to capture specific moments such as annotation, circling or highlighting words or phrases and answering items.

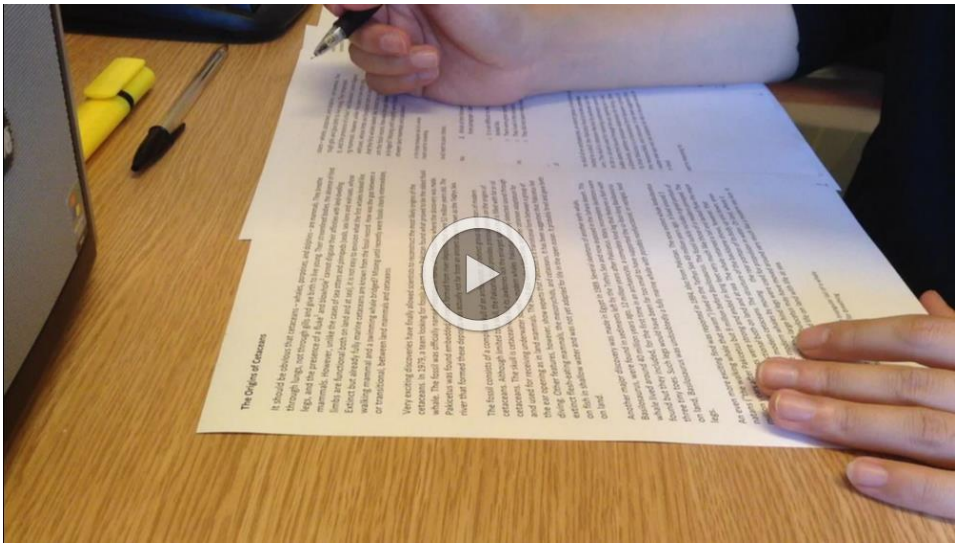


Figure 3.5. Video still of participant completing TOEFL reading paper (pilot study)

The participant answered ten items within 20 minutes, bypassing one, and not progressing to the final three items. They were thus discounted from further analysis. Participant responses to the items are outlined in table 3.15 below, recording five correct responses and 6 incorrect. As the participant was unfamiliar with the test format and item types, this success rate is unsurprising, emphasising the need to familiarise participants with the requirements of unfamiliar test tasks in the main study.

Question	Correct response	Participant response	Correct/ incorrect
1	B	A	x
2	A	C	x
3	C	C	✓
4	C	A	x
5	A	A	✓
6	B	C	x
7	D	D	✓
8	D	D	✓
9	B	--	N/A
10	C	C	✓
11	D	A	x

Table 3.15. Participant responses (pilot study)

Participant comments were more detailed with the use of video and allowed specific moments in the test to be targeted by the researcher and the participant, and provided a much more organic and immersive experience of test completion in place of the procedural description of item completion associated only with item responses in the initial studies. The participant was also able to verbalise without interviewer prompts. The participant was informed that the video could be paused for consideration, allowing the interview to be focused on moments of maximum participant recall. The participant paused the video on several occasions to explain what was occurring at those moments. Interviewer verbalisations remained consistent with those from initial studies two and three ('what were you thinking about at the time?') during moments of participant silence, and subsequent participant comments related to 'memory' and the use of past tense indicated that he was able to cast his mind to the moment of processing. Example verbalisations are included in table 3.16 below:

	Transcription
1	'… I bring my memory back to doing this kind of test, I remember we should read the question first, and this will be much more helpful to me…'
2	'I was imagining they would ask me something about …'
3	**15:52**. Underlines 'Ambocelatus swam like modern whales'
4	'…they may ask me how they swim, what they look like or something… because those sentences that I highlighted gave me some information rather than just passing [reading] the sentences'
5	[**PAUSES**] 'Let me think about what I thought'

Table 3.16. Excerpts of participant verbalisations from pilot study

The first example demonstrates that the participant is able to explain what he was thinking at that precise moment, detailing a specific item completion strategy to approach an unfamiliar paper. The second example demonstrates that the participant was able to further strategise whilst reading (monitoring their developing understanding of the text and anticipating what will come next) – tracing the text with a pencil and attempting to extract the salient *ideas* (not just key words) before reading the questions, confirmed with video evidence of goal-oriented action at 15 minutes and 52 seconds in the recording. Evidence of the success of this strategy is highlighted in example four, when the test taker correctly identified the main points of the text and was able to use this information to subsequently

answer question 10 correctly. This was coded as 'establishing propositional meaning at the sentence level'.

Example 5 explicitly shows that the participant is trying to remember their thoughts from that moment rather than reinterpreting and presenting them metacognitively. In SRI, verbalisations will always have a metacognitive element. The requirement of the task is for participants to *recall* their thoughts rather than replicate them exactly. Fidelity will always be partially compromised. Accepting this, the SRI must recreate the conditions of the test experience as closely as possible to maximise recall of thought processes at specific moments.

Also of note is the behaviour of the test taker in drawing on knowledge gained from previous sentences (excerpt 4, table 3.6). This data undermines a claim of *local independence* from a statistical perspective in these items, but provides evidence that test takers do not interpret items individually, but use the knowledge of the text that they have gained from previous items to either eliminate options in subsequent items; evidence that reading is constructively responsive (Pressley and Afflerbach, 1995) and that the methodology can capture moments of participants building meaning across sentences. Finally, the participant was also provided with verbal feedback on their performance as encouragement to view the testing and feedback session as part of ongoing language learning. This also encouraged a higher level of motivation for the participant to use the test to examine their own progress.

3.4. Summary

This chapter has outlined the methods of data collection and analysis that were carried out to address the four main research questions identified in the literature review. A series of twelve stimulated-recall interviews were conducted with six Chinese students, each of whom completed part of an IELTS and a TOEFL test. The tests used in the study were carefully selected to be as representative as possible. Interviews were conducted with three main stimuli; video, annotations and item responses. Verbalisations were carefully coded according to a schema devised from Khalifa and Weir's (2009) model of reading. These

codings were then quantified according to a rubric devised by Cohen and Upton (2006), allowing for comparison across tests and to identify the relative importance of strategies and processes for each item type. The findings, and how these relate to the wider literature will be presented in the next chapter in relation to each of the research questions in turn.

4. Findings and Discussion

4.1. Introduction

This chapter presents the findings from the stimulated-recall interviews in relation to the research questions identified in the literature review. The findings will be discussed in relation to each research question in turn. Research question 1 relates to the evidence model of the RE framework. As outlined in section 2.4 of the literature review, the evidence model details the evidence required to make inferences from observable variables (work products) to the construct of interest. This research question therefore identifies the observable task completion strategies that participants used to complete the IELTS and TOEFL tests. This resulted in the development of a strategy taxonomy which emerged organically form the data and applied directly to the test completion procedures of both IELTS and TOEFL. Task completion strategies used in the two tests are then compared directly.

Research questions 2 and 3 relate to the student model of the RE framework. The student model describes the construct in terms of what is being measured. The construct is interpreted in this study as the cognitive processes that the test takers employ in order to complete test tasks. Section 4.3 explicitly identifies the cognitive processes that the participants use in relation to task completion, interpreted via the coding scheme developed from Khalifa and Weir's (2009) model of reading and the decision-making algorithm used to infer relevant processes from participant verbalisations and observable behaviour.

Research question 3 specifically questions whether there are any differences between how participants complete the IELTS and TOEFL tests in terms of cognitive processing. This question therefore compares the overall use of specific cognitive processes in the two tests and compares the relative importance of each of the processes from Khalifa and Weir's (2009) model to the overall range of processes identified in relation to each test. This question addresses whether the test developers have similar global conceptions of the

construct of reading in English for academic purposes or whether there is a fundamental difference which undermines comparability.

Research question 4 relates to the task models of the RE framework. In critical parallel RE, task models consider whether individual item types in the two tests are associated with specific cognitive processes, and therefore specific claims made about successful test takers. Where research questions 2 and 3 address the data at the test level, research question 4 uses the comparative rubric developed by Cohen and Upton (2006) and employed in this study. Using the rubric will demonstrate the relative importance of specific cognitive processes and whether they can be identified with specific item types, or whether the test developers intend that individual items target a range of cognitive processes.

The final section of this chapter will summarise what has been learned from the development of the framework of RE and to what extent the development of this framework has brought together aspects of enquiry which otherwise would not have been possible. This section will also consider the utility and application of the framework to wider test development concerns, and consider the methodological innovations which resulted from the application of this framework.

4.2. Research question 1: What observable test-taking strategies do test takers use when completing IELTS and TOEFL reading tests? Are there any differences in how participants respond to IELTS and TOEFL tests?

This section directly answers research question 1 by identifying the test-taking strategies that the research participants used in completing both IELTS and the TOEFL tests. The principal means of identification was video evidence. Actions that were visible on-screen were described and added to a strategy rubric. Each individual observed instance of this action was then coded and counted. As new actions were observed, they were added to the rubric. This procedure occurred with all video evidence of participants engaging with both IELTS and TOEFL. The outcome is a strategy rubric which covers all strategic actions and allows direct comparison between the two tests. Not all strategies are directly observable

form only video evidence. Some strategies were confirmed based on participant verbalisations which complements video evidence. These are identified as such in the discussion below. The first section clearly identifies the strategies which were observed from video and participant verbalisations. These are associated with the outer, metacognitive columns of Khalifa and Weir's reading model. Frequencies of individual strategic actions are then identified for the two tests. Sections 4.2.2 – 4.2.4 then specifically outlines similarities and differences between IELTS and TOEFL. Section 4.2.5 considers whether specific strategies are related to particular item types. Section 4.2.6 then discusses the findings in relation to existing literature on strategies identified in IELTS and TOEFL, and argues that the approach to identifying strategies in the present study has provided far more finely-grained information on strategic actions, and therefore a greater *evidence model* from which to infer underlying cognitive processes.

4.2.1. Identification of test takers' strategic actions used to complete IELTS and TOEFL

Thirty-four individual strategic actions were identified from the video evidence. The schema emerged organically from analysis of the videos, thus is a largely chronological representation of participant progress through the test, designed to elucidate the developing understanding of the text and items. Codes 1-10 detail participant engagement with the items, item options, and text heading. Codes 11-17 are sub-categories of 'careful reading' and provide analytical scope for understanding how a participant engages with the text. Codes 19-24 detail expeditious reading behaviour when the participant is attempting to gain an overview of the text or to locate specific information by scanning the text. The remaining codes (18 and 25-34) are specific test management strategies which provide information regarding how participants engage with items and text simultaneously and attempt to integrate information. These codes relate broadly to *monitoring* activities. The strategies and the respective frequencies of each for all six participants are displayed in table 4.1 below:

Code	Descriptor	IELTS	TOEFL	Total
1	Reads question(s) before proceeding to the text	12	37	49
2	Identifies the purpose of the question	13	61	74
3	Reads text heading	2	4	6
4	Identifies grammatical or content-based parallel between items/options	5	1	6
5	Reads question stem(s) and/or option(s) carefully	9	62	71
6	Marks/notes key noun phrase(s) in the questions stem or options	75	59	134
7	Marks/notes key verb phrase(s) in the question stem or options	19	10	29
8	Marks/notes key adjective phrase(s) in the question stem or options	17	21	38
9	Marks/notes key prepositional phrase(s) in the question stem or options	6	2	8
10	Marks/notes key adverbial phrase(s) in the question stem or options	2	6	8
11	Careful local reading (text)	72	57	129
12	Marks/notes key noun phrase in the text during careful reading	42	113	155
13	Marks/notes key verb phrase in the text during careful reading	23	26	49
14	Marks/notes key adjective phrase in the text during careful reading	15	13	28
15	Marks/notes key prepositional phrase in the text during careful reading	7	3	10
16	Marks/notes key adverbial phrase in the text during careful reading	0	2	2
17	Marks/notes key conjunction in the text during careful reading	0	1	1
18	Returns to the question for clarification: rereads question and/or options	49	29	78
19	Searches for key word/phrase (text)	25	6	31
20	Skimming part of the text for general understanding (expeditious reading)	7	15	22
21	Marks/notes key phrase in the text during expeditious reading	5	11	16
22	Identifies content-based parallel between paragraphs	7	8	15
23	Identifies lexical parallel between parts of the text (matches words across paragraphs or heading)	4	2	6
24	Identifies paraphrase within text or between text and item stem	14	12	26
25	Eliminates option(s) (contradicts key noun/verb phrase/nonsensical response)	10	34	44
26	Eliminates option (s) (no information found)	7	13	20
27	Compares question stem/option to a portion of the text	38	32	70
28	Checks/confirms/considers option choice after reading portion of text	19	45	64
29	Guessing	3	4	7
30	Hesitates while answering to reconsider choice	10	9	19
31	Checks/confirms response has met the parameters of the task	3	10	13
32	Uses own topic knowledge to enhance understanding of text/questions	4	4	8
33	Writes note/labels part of text/item	17	11	28
34	Checks progress	2	3	5

	Total	**548**	**728**	**1276**

Table 4.1. Participants' use of strategies in IELTS and TOEFL

Each represents a specific action that was directly observable as the participants completed both the IELTS and TOEFL tests, with the exception of codes 22, 24 – 26, 29, 31 and 32, which rely on participant verbalisation to code in addition to observable actions. For example, a participant may be observed to read more than one part of a text carefully and move between these two parts of the text, but interview verbalisations are required to state that the participant is successfully comparing text content or that he or she has eliminated an option because it does not cohere with the part of the text they were previously reading. Similarly, when a participant responds to an item, verbalisations are specifically required to be certain that the participant is guessing a particular option. Verbalisations are also required to know that the participant has specifically used their own topic knowledge to assist them with an item or text comprehension. These codes were nonetheless categorised as 'strategic decision-making because they form part of the test-takers' metacognitive repertoire of skills used for test management. Table 4.2 below details how the emergent coding scheme relates to the right and left-hand sides of Khalifa and Weir's model of reading:

Metacognitive Activity		**Monitor**	
Remediation where necessary	30	**Text structure knowledge:**	
Monitor: goal checking	1, 2, 18, 31, 33, 34	Genre Rhetorical tasks	3 4, 27, 28
Goal setter Selecting appropriate type of reading:		**General knowledge of the world**	32
Careful reading **LOCAL** Understanding sentence	3, 5, 11-17, 24, 25, 27, 28	**Topic knowledge**	3, 32
GLOBAL Comprehending main idea(s) Comprehend overall text(s)	22, 26,	**Meaning representation of text(s) so far**	22, 26, 28, 33

Expeditious reading **LOCAL** Scan/search for specifics	4-10, 19-21	**Syntactic knowledge**	5, 11, 24, 25
		Lexicon Lemma: Meaning Word class	6-10 12-19, 21, 23
GLOBAL Skim for gist Search for main ideas and important detail	23, 24	**Lexicon Form**: Orthography Phonology Morphology	

Table 4.2. Strategy coding scheme mapped to Khalifa and Weir (2009) metacognitive and monitoring components

All of the identified strategies were able to be categorised into one of the two columns. These strategies provide strong empirical evidence for the outer columns of the model. In particular cases, such as 'monitoring' and 'careful local reading', a number of individual sub-strategies were identified, suggesting that these strategies are too complex to be regarded as single strategies. 'Careful local reading' (code 11) refers to instances in video recordings in which participants are clearly engaging carefully with part of the text, without making any other discernible actions. Codes 12-17, 24-25 and 27-28 represent actions taken as a result of engaging closely with a particular portion of the text. Others, such as genre identification, only have a single strategic action evident in the rubric (code 3), which was confirmed by participant statements that they inferred information about the type of text with which they were about to engage based on information contained within the heading. Having identified 34 explicit strategic actions, the next sections focus on similarities and differences in strategic actions that were evident in participants' engagement with IELTS and TOEFL. Sections 4.2.2 and 4.2.3 identify explicit differences, which are discussed in section 4.2.4.

4.2.2. Observable differences exist in how participants respond to IELTS and TOEFL tests in terms of strategic actions

The video approach to the stimulated-recall methodology resulted in the identification of 1276 individual moments of participant strategic decision-making for the six participants, an average of 106.33 for each testlet, demonstrating the efficacy of a finely-grained approach to observable strategy identification. Participants recorded a higher frequency of observable strategic actions when taking the TOEFL test than the IELTS test (57 per cent of all observed

strategies related to TOEFL). Graph 4.1 (representing the raw data in table 4.1) below detail the strategies used by participants on the IELTS and TOEFL test. The figures reveal some striking differences between how participants approach task completion in the two tests.

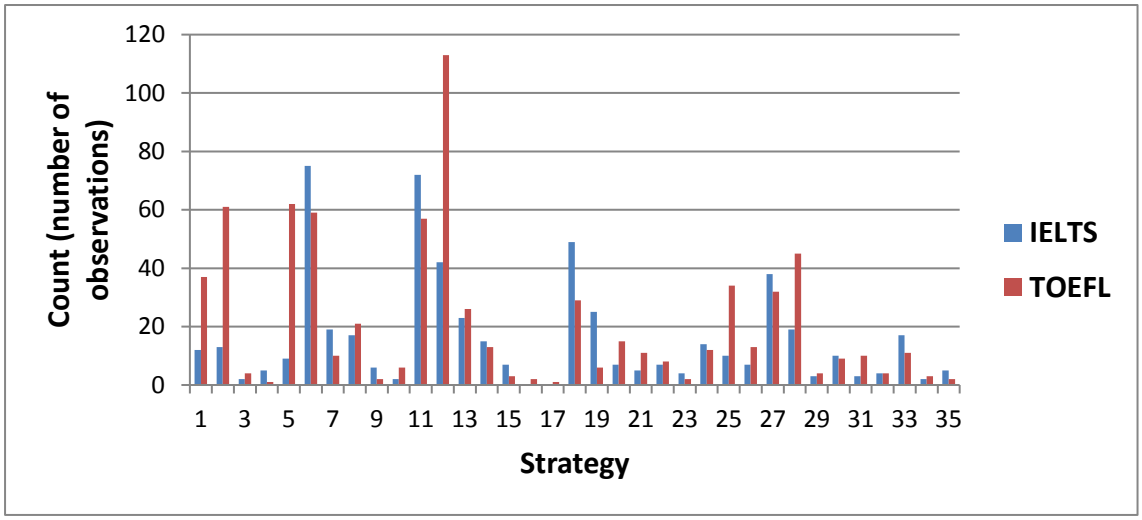


Figure 4.1. Number of observations for each strategy for IELTS and TOEFL

Figure 4.1 displays the total number of strategies observed for all six participants for both the IELTS and TOEFL tests. Code 35 refers to identified instances of test-wiseness, for which five instances emerged in relation to IELTS and two in relation to TOEFL for the six participants. The five most frequently observed strategies (raw data) for each test are listed in table 4.3 below:

IELTS top 5 reported strategies		Freq.	TOEFL top 5 reported strategies		Freq.
6	Marks/notes key noun phrase(s) in the questions stem or options	**75**	**12**	Marks/notes key noun phrase in the text during careful reading	**113**
11	Careful local reading (text)	**72**	**5**	Reads question stem(s) and/or option(s) carefully	**62**
18	Returns to the question for clarification: rereads question and/or options	**49**	**2**	Identifies the purpose of the question	**61**
12	Marks/notes key noun phrase in the text during careful reading	**42**	**6**	Marks/notes key noun phrase(s) in the questions stem or options	**59**
27	Compares question stem/option to a portion of the text	**38**	**11**	Careful local reading (text)	**57**

Table 4.3. Top five reported strategies for IELTS and TOEFL

Some clear similarities and differences emerged in terms of most commonly deployed strategies. In both tests, participants underlined key noun phrases as focal points for further careful reading. However, in IELTS, participants performed this more frequently in item stems than in the text itself, indicative of greater complexity in the items, as there is more information to process. IELTS item types such as 'matching headings' require participants to revisit the item on multiple occasions as there is a substantial amount of information they must process in order to eliminate incorrect options and to match the key to portions of the text. The frequency in which participants return to the question and compare parts of the question to the text indicate the extent to which this item type influences test taker behaviour in IELTS. By contrast, in TOEFL, participants are careful to initially identify the purpose of the question and identify key nouns, but then are much less likely to return to the question for clarification.

4.2.3. Commonalities in strategies employed by test takers to IELTS and TOEFL reading

The most frequently-observed behaviour by the participants in all sessions was underlining (or highlighting by other means) key nouns in the text and question stems (strategies 6 and 12). For IELTS, this process occurred most frequently in relation to the question stem, and for TOEFL, the text. The identification of key nouns was used as a search strategy by all participants as a means of locating the relevant portion of the text. Despite the TOEFL test reproducing the paragraph that relates to each question, participants were aware that specific items will still relate to one or two target sentences in the text and will therefore use the key words to orient themselves. Strategies 6, 11 and 12 occur in the top five most frequent strategies for both tests. This creates a broadly consistent general picture of item completion for participants in both tests. Participants tended to highlight key noun phrases in the question stem and then proceed to the text to identify those phrases, then read carefully around them. From a processing perspective, this suggests that word matching/lexical access was a common approach to answering items. Therefore, a high frequency of *lexical access* process coding can be anticipated.

In relation to the TOEFL test, participants then tended to identify and mark further noun phrases that relate to the question stem or options. Participants underlined a much greater proportion of noun phrases in the TOEFL test compared to IELTS. This pattern did not continue for other parts of speech, indicating the relative importance of noun phrases to item completion in TOEFL. This can be linked to research question 5; is there a statistically significant difference between the texts used in IOELTS and TOEFL in terms of parts of speech and how did these differences feed into item development.

4.2.4. Differences in strategies employed by test takers to IELTS and TOEFL reading

For IELTS, code 27 occurred in the most frequent strategies, but did not for TOEFL. Code 27 is used when participants are directly comparing a part of the text to the question stem or option. The frequency of this strategy for IELTS in relation to TOEFL may be due to test method effect. For example, IELTS test 2 (items 14-18) contains a *matching headings* series of items. This item type requires multiple revisits due to the high number of options available to test takers, which they would be unable to hold in their working memory. This could also be an artefact of the research method. Test papers were reproduced as accurately as possible from source material. For TOEFL, items appear on the screen next to the text. For IELTS, they are on a different page. This manifested in the research as questions and text appearing on different pages for IELTS items in comparison to TOEFL items, in which items appear beneath the relevant paragraph. Thus, revisits may be more visible for IELTS than TOEFL in the video stimulus. One advantage of eye-tracking studies that the current study cannot duplicate is the determination of when an individual is looking at an item or the text, as the direction of the video is on the test paper and participants' hands, rather than their eye movement. Nonetheless, TOEFL tests had greater strategic action overall, possibly due to lack of familiarity with the item types.

For IELTS, code 18 occurs in the most frequent five strategies, but does not for TOEFL. Code 18 refers to instances in which the participant returns to the question for clarification (monitoring/goal checking). This differs from code 27 as code 18 refers to instances in which the participant returns to read the question stem or directions rather than the options.

TOEFL items follow a more standardised format than IELTS items and all are presented to the participant in the same, four-option multiple choice format, prompting fewer revisits to the question stem. At the same time, code 5 appears in the most frequent strategies for TOEFL but not for IELTS. Participants noticeably spent longer reading item stems. Although TOEFL offers multiple examples of the same item type in the test, these may not always be located concurrently in the test, necessitating careful reading of the stem and/or directions for each item. For IELTS, items are grouped per item type, with one set of instructions for each. This would also explain the time spent by participants initially reading the stem to familiarise themselves with the requirements of the task. Code 2 is more frequently observed for TOEFL than for IELTS. Code 2 indicates the participant purposefully spends time identifying the requirements of the item type before moving to the question stem to identify the information needed to successfully complete the task. This could also be due to the lack of familiarity of the participants with the item types and the arrangement of item types in the two tests. The difference in test format would also explain why participants are observed to read the question in TOEFL before proceeding to the relevant text, whereas for IELTS participants tended to answer examples of common item types concurrently. Items in IELTS are not fully locally independent, as the selection of one heading in a '*matching headings'* series of items will influence the probability of success in the subsequent item that the participant attempts.

These strategy differences resulted in a more intermittent engagement with the text in IELTS tests than in TOEFL tests. Each instance of careful local reading (code 11) was coded as participants moved to the text and engaged with a specific part of it. The codings reflect each instance of a participant making a decision to engage with part of the text; they do not reflect the total amount of time engaging with the text. The level of engagement with the text is evidenced from verbalisations in participants' explanation of their actions, which was subsequently coded in relation to the processing core.

Strategies 25 (eliminates option(s); contradicts key noun/verb phrase/nonsensical response) and 28 (checks/confirms/considers option choice after reading portion of text) were more frequently observed for TOEFL than IELTS. The difference in code 25 between the tests is likely method effect as the greater number of multiple-choice items would lead to an

elimination strategy that is less likely in relation to sentence completion, identifying information, identifying writer's view items. The difference in code 28 is less likely to be due to method effect. It suggests that participants were less certain of their responses to TOEFL items than IELTS items, which could be due to more complex processing associated with those items that participants are more likely to check.

4.2.5. Strategic differences by item type in IELTS and TOEFL

As stated in the methodology section 3.3.4.3, data across tests would be compared using a modified version of the metric devised by Cohen and Upton (2006). This would provide a basis for comparing across item types for each of the tests, comparing across tests and comparing the findings related to the TOEFL test directly to the findings of Cohn and Upton (2006). The number of strategies identified for each item type was divided by the number of items representing that item type to get a perspective of the relative importance of each of the strategies for that item type.

Frequencies of strategy usage were broken down per item type, and then divided by the number of items in that type. The resulting figures were then categorised according to the schema in table 4.8. One weakness in the current study that needed to be addressed in replicating Cohen and Upton's methodology was the reliance on a single example of each test. A complete TOEFL test contains only three reading-to-learn items, only one of which was a schematic table. Therefore, in analysing the strategic actions of participants, these were combined to mitigate any extreme values from distorting the findings. This resulted in eight individual item types for both IELTS and TOEFL with values for each of the 34 identified strategies. This created a matrix of 272 individual cells for each test. This produced the following tables, with darker shading indicating higher frequency. Table 4.4 records the frequency of categories for each test. Table 4.5 and Figure 4.2 display the proportions of strategy use across the two tests:

	IELTS								TOEFL							
Strategy	1-MC	2-IDi	2-ID	3-IDw	4-MI	5-MH	8-SC	9-SuC	BC-v	BC-(n)f	BC-ss	BC-pr	I-rp	I	I-it	R2L-ps; R2L-st
1	L	L	S	S	S	L	L	M	M	H	H	S	H	M	H	H
2	H	S	S	S	L	L	M	S	VH	H	H	H	VH	VH	H	H
3	S	S	S	S	S	S	S	S	S	S	S	S	S	S	S	S
4	M	S	S	S	S	S	S	S	S	S	S	S	S	S	L	S
5	H	S	S	S	M	L	S	S	H	H	H	M	VH	VH	VH	H
6	VH	H	H	H	VH	L	M	H	L	VH	S	S	H	VH	VH	VH
7	L	S	H	S	H	S	S	S	S	M	L	S	S	M	S	S
8	S	L	M	S	H	S	S	S	S	M	H	S	L	H	L	M
9	H	L	S	S	L	S	S	S	S	S	S	S	S	S	L	S
10	L	S	S	S	S	S	S	S	S	S	L	S	L	S	S	L
11	VH	VH	M	VH	H	H	VH	H	H	H	M	S	H	VH	H	VH
12	VH	H	L	S	H	L	H	H	M	VH	VH	H	H	VH	S	VH
13	VH	H	M	S	M	S	L	M	M	M	S	S	L	H	L	M
14	M	S	S	S	M	L	S	S	S	S	L	S	L	S	L	L
15	H	S	S	S	S	S	S	S	S	S	S	S	S	S	S	L
16	S	S	S	S	S	S	S	S	S	S	S	S	S	S	S	L
17	S	S	S	S	S	S	S	S	S	S	S	S	S	S	S	S
18	VH	VH	M	M	H	H	H	H	S	H	M	S	VH	S	L	H
19	H	L	S	M	M	L	H	H	S	S	S	S	S	S	S	L
20	M	L	S	S	S	L	S	S	L	S	S	M	S	M	S	M
21	M	S	S	S	S	S	S	M	S	S	S	S	S	S	S	S
22	H	S	S	S	S	L	L	S	S	S	S	S	L	S	S	S
23	L	S	S	S	S	L	S	S	S	S	S	S	S	S	L	S
24	M	S	L	M	H	L	L	S	S	M	H	S	L	S	L	L
25	VH	S	S	S	L	S	S	S	M	M	H	M	M	VH	H	VH
26	M	S	S	S	M	S	S	S	S	L	H	S	M	S	S	M
27	H	M	M	VH	H	H	M	H	M	H	H	S	H	H	L	H
28	L	S	M	H	L	M	L	M	S	M	VH	H	M	VH	H	VH
29	S	S	S	S	L	S	S	S	S	S	S	S	L	S	S	S
30	S	L	S	S	M	M	S	S	S	S	S	H	L	S	L	M
31	S	S	S	S	S	L	S	S	S	S	S	S	M	S	M	M
32	S	S	S	S	S	L	S	S	S	S	S	S	L	M	S	S
33	S	S	H	S	VH	S	S	S	S	S	S	S	S	S	L	VH
34	S	S	S	S	S	S	S	S	S	S	L	S	S	S	S	S

Table 4.4. Weighted frequency of observed reading and test-taking strategies for IELTS and TOEFL

Frequency	IELTS	TOEFL
Very High	13	23
High	34	38
Moderate	31	31
Low	36	33
Sporadic	158	147
Total	**272**	**272**

Table 4.5. Number of each category for IELTS and TOEFL

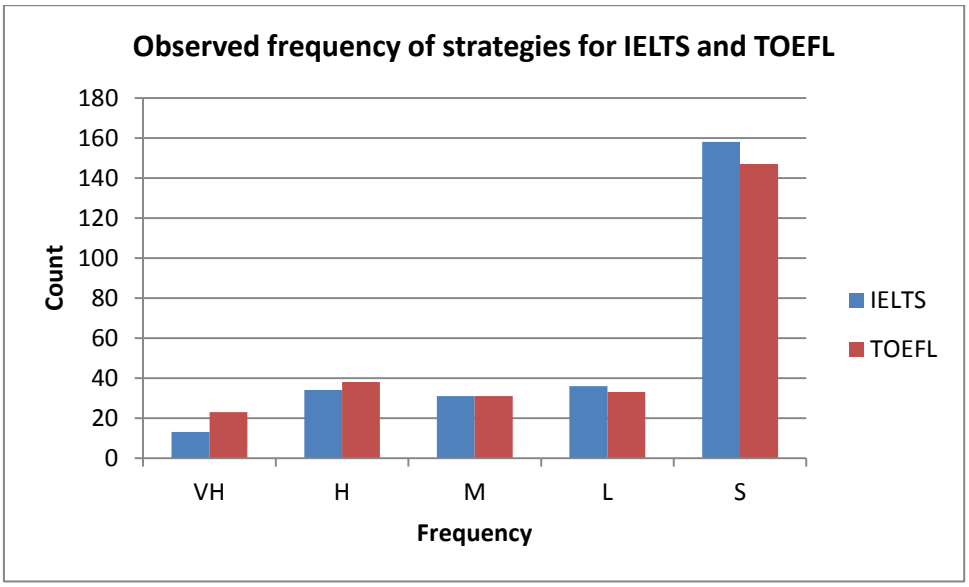


Figure 4.2. Observed frequency of strategies for IELTS and TOEFL

The data reflects the general observation that most strategies were used sporadically across all item types, and that particular strategies that recorded 'very high' of 'high' ratings for one item were more likely to record similar ratings for other item types. Participants therefore relied upon a particular set of strategies that they were well-rehearsed in for test completion tasks. The data showing the top five strategies for both tests identified 6, 11 and 12 as occurring for both IELTS and TOEFL. This data reveals how these strategies were used across the item types. For IELTS, strategy 6 is more consistent across item types than TOEFL. For TOEFL, marking key nouns in the question stem is a strategy used more frequently for item types that target higher-level processing, with the exception of basic comprehension, negative factual items, which requires participants to assess each option carefully. For IELTS, this strategy was rated 'low' for matching heading items. This is likely due to the shorter length of the options, allowing participants to hold key information in their working memories.

Participants were more likely to read the questions before proceeding to the text in TOEFL than in IELTS. This is true for all item types, with the exception of pronoun reference item types in TOEFL. Once participants had familiarised themselves with this item, they were more inclined to move to the text and complete the item after they had familiarised themselves with the text. Participants also spent longer familiarising themselves with task requirements in TOEFL. This is expected as they were more familiar with IELTS than TOEFL item types. This was universally true for all item types in TOEFL, although participants carefully read requirements of the multiple-choice item types in IELTS. Strategy 5 (Reads question stem(s) and/or option(s) carefully) was a noticeable difference between IELTS and TOEFL. This strategy is widely used across item types in TOEFL. Only one item type (multiple-choice) was rated as 'high' for this strategy in IELTS. As most items in TOEFL are multiple-choice, this is further evidence that this strategy was highly visible as an artefact of test method. Strategy 11 (careful local reading) was used by all participants in response to all item types in the two tests with the exception of pronoun reference in TOELF tests. Participants read the text initially and were able to answer this item type often without further reference back to the text. Strategy 25 (Eliminates option(s), contradicts key noun/verb phrase/nonsensical response) also displayed significant differences between IELTS and TOEFL. As this was almost exclusively used in relation to IELTS multiple choice items and more generally in TOEFL, this strategy can also be said to be an artefact of task design associated with multiple choice items. Strategy 18 (returns to the question for clarification: rereads question and/or options) was used by participants more frequently in relation to IELTS than TOEFL. With the data for strategy 5, this suggests participants were inclined to read TOEFL items carefully the first time they encounter them, prompting fewer revisits. With IELTS, participants did not read items carefully initially, but then were more likely to revisit them as their understanding of the text grew. Revisiting items in IELTS was broadly consistent across item types. For TOEFL, this strategy was most noticeable for inferencing (rhetorical purpose) item types, suggesting that this item type placed particularly large cognitive demands on the test takers.

The data also reflects the general trend that strategy use was more prevalent amongst participants when they were completing the TOEFL test than the IELTS test. Introducing an

additional category allowed more differences to emerge between IELTS and TOEFL to emerge in this categorical data. The TOEFL test recorded a greater proportion of VH codes (23) than IELTS (13), which correspondingly saw more 'sporadic' use of strategies. Numbers of 'high', 'moderate' and 'low' strategy use was broadly consistent across the two tests. Seven strategies recorded sporadic or low ratings across all item types for both tests:

3	Reads text heading
10	Marks/notes key adverbial phrase(s) in the question stem or options
16	Marks/notes key adverbial phrase in the text during careful reading
17	Marks/notes key conjunction in the text during careful reading
29	Guessing
23	Identifies lexical parallel between parts of the text (matches words across paragraphs or heading)
34	Checks progress

Table 4.6. Sporadic/infrequently-cited strategies across both tests

The grounded nature of the methodology of the research meant that inevitably strategies would emerge that would be infrequently cited. Infrequent use of these strategies suggests they may be of less utility to future scholars, who could make decisions about whether or not to include them based on their own requirements and analysis of test instruments to consider whether their inclusion would be fruitful. Conjunctions and adverbial phrases are unlikely to be highlighted by participants unless specific items are targeting that vocabulary. Infrequent use of guessing (code 29) as a strategy is desirable as widespread guessing undermines the aims of the test. Only seven instances of guessing were cited by the participants in the research, although only after participants had attempted alternative methods of getting an answer. Strategy 23 is useful in attempting to observe participants form mental models or establish a coherent picture of a text. This activity is unlikely unless an item requires participants to build knowledge across paragraphs as a requirement of an item. Strategy 34 refers to instances in which participants review their progress by examining the test itself – how many items they have completed and how many they still have to go. This does not relate directly to construct of reading, but was coded as an observable example of test management behaviour.

4.2.6. Comparison of findings with literature relating to test strategies used in IELTS and TOEFL

This section compares and discusses the strategic findings of the study with two studies that examined strategic test management in IELTS and TOEFL which were highlighted in the literature review. This section specifically focuses on the findings in relation to the strategic findings of Weir et al (2009a) and Cohen and Upton (2006), as these two studies present the most comprehensive evidence of strategic management of the two tests. The final part of this section then broadens the discussion to consider the wider strategic literature, and how the division of strategies and processes in the literature informed the study to produce more convincing findings. Weir et al (2009a) investigated strategic processes intrinsic to the academic reading component of the IELTS, using retrospective questionnaires, verbal reports and expert judgement. Cohen and Upton (2006) present the most comprehensive account of strategic management of the reading component of the TOEFL test, also based on verbal report data. These studies are therefore very important for comparing and informing the findings of the current study. Data from these studies presented in the literature review is reproduced in this section for ease of comparability. This section also considers how the findings link to current thinking in relation to metacognition and monitoring.

Weir et al (2009a) identified a range of careful and expeditious reading strategies based on the Khalifa and Weir componential model of reading. A matrix containing these strategies cross-referenced against item types was provided to two participants (A and B) who completed the matrix independently on fourteen IELTS tests. Table 2.4 in the literature review is reproduced below to assist in interpretation of the findings from the present study:

	Read expeditiously by	**Read carefully for meaning which is**	**Reading**	

Test taker analyst	skimming	Search reading	scanning	Explicit within sentences	Implicit within sentences	Explicit across sentences	Implicit between sentences	To construct a text model	For a situation model of text and own prior knowledge	Totals per reader analyst
A	0	45	50	277	27	115	45	3	0	**562**
B	70	6	93	318	12	57	25	4	0	**585**
Cognitive skills total	**70**	**51**	**143**	**595**	**39**	**172**	**70**	**7**	**0**	**1154**
Sub-totals: expeditious vs careful reading	**264**			**883**						**1154**

Table 4.7. Summary of responses to two test taker analysts to reading test tasks of 14 authentic IELTS reading tests (Weir et al, 2009a).

The findings of interest in this section are the subtitles of 'reading expeditiously' (and sub-skills of scanning, searching and skimming) and 'reading carefully'. The subskills for the latter category are classified in the current study as 'processing' rather than strategies, so will be of interest in relation to the research questions addressing processing. Table 4.7 presents data for fourteen IELTS tests (560 items in total). Comparative data showing the cumulative strategic findings in the current study (two tests, 78 items for six participants) are presented in table 4.8 below for comparative purposes:

	Careful reading	**Expeditious reading**	**Total**
IELTS	265	260	525
TOEFL	425	135	560
Total	690	395	1085

Table 4.8. Careful and expeditious reading in IELTS and TOEFL from data in the present study.

Table 4.8 presents cumulative data from table 4.2 containing codes relating to careful and expeditious reading. Fifteen codes from the coding scheme in table 4.2 align with the definition of 'careful reading' offered by Khalifa and Weir (codes 3, 5, 11-17, 22, 24-28). Twelve codes align with 'expeditious reading' (codes 4-10, 19-21, and 23-24). 690 of the 1085 total codes comprise expeditious reading (63.6%), less than the proportion identified in Weir et al (2009a). However, of the total number of careful reading codes, 425 (61.6%) are linked to TOEFL rather than IELTS, suggesting that participants spent proportionally

more time reading the texts carefully in TOEFL than they did with IELTS. The total number of codes identified for IELTS was split almost evenly between careful and expeditious reading. Of the expeditious reading codes, 260 out of 395 (65.8%) relate to IELTS. Just over fifty percent (50.5%) of the codes for IELTS were linked to careful reading, in contrast to the findings of Weir et al (2009a), whose participants reported more than three times as many instances of reading carefully as reading expeditiously. 883 of 1154 reported strategies (76.5%) were instances of participants reading carefully rather than expeditiously (Weir et al, 2009a).

Participants in Weir et al (2009a) also reported no engagement with the text as a whole. Patterns of expeditious reading clearly differed across the two participants, although the authors indicate that this discrepancy may be due to overlap between individual strategies which resulted in participants reporting different strategies. Analyst A also skim read the passages before responding to the items, so constant search reading through the texts once she had gained a picture of the topics and structure was not necessary. In contrast, the strategies in the current study were individually identified based on video observation and were defined by the researcher. Each observed action was coded independently of others, meaning there was no duplication (individual strategic actions coded more than once).

There are clear similarities and differences between the findings in this study and the findings of Weir et al (2009a) regarding how test takers approach IELTS. Weir et al expected the research to demonstrate that each item would elicit a single skill. Weir et al's study (2009a) assigned two participants to analyse the tests, and based their findings on whether the participants reported the use of specific strategies for individual item types. Due to this methodology, the authors did not report finely-grained analysis for each of the item types. However, the higher totals indicate that for some items, the participants reported more than one skill for each item. Data in the current research backs up this assertion: a total of 1085 individual strategic moments relating to either careful or expeditious reading were identified relating to a total of 78 items. No single item type in either IELTS or TOEFL required the use of an individual strategy for successful completion. This indicates two significant outcomes. Firstly, this is strong evidence that test takers use a combination of strategies to complete each item, and that it is preferable to speak of relative importance in

relation to strategic management of individual test items, rather than claiming an item requires one specific strategy for successful completion. These findings hold for both IELTS and TOEFL.

There are also two problems with the authors' (2009a) study which this study has attempted to overcome. The first is reliance upon test taker metacognitive reflection rather than researcher analysis of participant verbalisations to determine which element of Khalifa and Weir's model the participant has activated. The taxonomy is presented in simple language, avoiding metalinguisic pitfalls (Field, 2004: 318). Test takers are capable of explaining their thought processes, but requiring them to recall whether specific responses were the product of engagement with a single sentence or more than one and the extent to which a response is the product of their schematic knowledge or knowledge of vocabulary or grammar may lead to confused responses. More abstract options are less likely to receive responses. This procedure may be exacerbated by the time delay (the second problem). Participants were asked to complete the strategy questionnaire after completing the test. This would result in a delay of fifteen to twenty minutes' delay between answering the first items and responding to the questionnaire related to those items.

Similar to the present study, Cohen and Upton (2006) used video to assist in the identification of strategies. This enabled the development of a more comprehensive strategic framework, similar to that produced in this study. Cohen and Upton's research design was clearly designed to capture multiple strategic actions by individual test takers in relation to individual items. The authors produced a strategy matrix using the ratio of strategies to each item type, adopting four categories, which was expanded to five in this study. The findings are reproduced from the literature review in tables 4.8 and 4.9 for comparison purposes:

Strategy	BC-v	BC-pr	BC-ss	BC-f	BC-n/e	I	I-rp	I-it	R2L-ps	R2L-st
R6	H	VH	VH	VH	VH	VH	VH	VH	H	VH
R7	L	M	L	L	L	L	M	H	H	H
R9	M	H	H	VH	H	H	H	H	M	VH
R10	L	L	L	M	L	L	L	L	L	L
R26	L	H	L	L	L	L	L	L	L	L
T1	M	M	M	H	H	H	H	M	M	H
T2	L	L	M	H	M	H	M	M	M	VH
T3	L	L	L	L	L	M	L	L	L	L
T4	L	L	L	L	L	L	L	L	H	H
T5	VH	VH	VH	VH	VH	VH	VH	VH	L	L
T6	L	L	L	L	L	L	M	L	L	L
T8	L	L	L	L	L	L	L	M	L	L
T10	H	M	L	L	L	L	L	L	L	L
T12	H	L	H	H	H	H	H	L	L	L
T13	H	L	L	L	L	L	L	L	L	L
T14	L	L	M	M	M	H	M	L	H	L
T16	H	H	VH	VH	VH	VH	VH	L	VH	VH
T17	L	L	M	L	L	L	M	L	M	H
T19	L	L	M	L	M	L	L	L	H	L
T21	H	L	L	L	L	L	L	L	L	L
T22	L	VH	H	H	H	H	H	H	VH	VH
T24	L	L	M	M	M	M	L	L	L	L
T26	L	L	L	L	L	L	L	M	L	L
T27	M	L	L	L	L	L	L	L	L	L
T28	VH	VH	VH	VH	VH	VH	VH	H	VH	VH

*Frequency rate = no. of occurrences / no. of items of that type. Rates ≥1.0 (marked VH) were classified as 'very high', rates ≥.50 (marked H) were classified as 'high', rates ≥.30 (marked M) were classified as 'moderate', and rates ≤.30 (marked L) were classified as 'low'. Strategies that were used at a low (≤.30) rate across *all* item types were not included in the table.

Table 4.8. Frequency of reported use of reading and test-taking strategies (Cohen and Upton, 2006: 225).

Approaches to reading the passage (R)

R6	Reads a portion of the passage carefully.
R7	Reads a portion of the passage rapidly looking for specific information.
R9	Repeats, paraphrases, or translates words, phrases, or sentences—or summarizes paragraphs/passage—to aid or improve understanding.
R10	Identifies an unknown word or phrase.

Inferences

R26	Verifies the referent of a pronoun.

Test-Management Strategies Coding Rubric (T)

T1	Goes back to the question for clarification: Rereads the question.
T2	Goes back to the question for clarification: Paraphrases (or confirms) the question or task.
T3	Goes back to the question for clarification: Wrestles with the question intent.
T4	Reads the question and considers the options before going back to the passage/portion.
T5	Reads the question and then reads the passage/portion to look for clues to the answer, either before or while considering options.
T6	Predicts or produces own answer after reading the portion of the text referred to by the question.
T8	Predicts or produces own answer after reading questions that require text insertion (I-it types).
T10	Considers the options and checks the vocabulary option in context.
T12	Considers the options and selects preliminary option(s) (lack of certainty indicated).
T13	Considers the options and defines the vocabulary option.
T14	Considers the options and paraphrases the meaning.
T16	Considers the options and postpones consideration of the option.
T17	Considers the options and wrestles with the option meaning.
T19	Reconsiders or double-checks the response.
T21	Selects options through background knowledge.
T22	Selects options through vocabulary, sentence, paragraph, or passage overall meaning (depending on item type).
T23	Selects options through elimination of other option(s) as unreasonable based on background knowledge.
T24	Selects options through elimination of other option(s) as unreasonable based on paragraph/overall passage meaning.
T26	Selects options through their discourse structure.
T27	Discards option(s) based on background knowledge.
T28	Discards option(s) based on vocabulary, sentence, paragraph, or passage overall meaning as well as discourse structure.

Table 4.9. Reading strategies which recorded a rating of 'moderate' or 'higher' in relation to at least one item type (Cohen and Upton, 2006: 220-222).

Table 4.9 represents a subset of strategies in the original rubric used by Cohen and Upton (2006). Their strategy rubric contains fifty-nine codes, of which twenty-six recorded 'medium' or higher use in relation to at least one item type. The first clear difference between the rubrics is the number of codes in the present study that relate directly to 'careful reading', which in Cohen and Upton's study is composed of a single code (R6). 'Reading carefully' has been expanded upon in the rubric in this study. Test takers spent

considerable time reading carefully, and undertook a number of directly observable actions while reading carefully. Coding this behaviour allows for more careful consideration of participant engagement with the text and more evidence to carefully link strategic actions with underlying processing.

Of the strategies which emerged from the present study (table 4.1), seven (3, 10, 16, 17, 23, 29, 34) out of thirty-four recorded only 'sporadic' or 'low' use across item types for both tests. This indicates that a video-based approach to strategy identification is successful in identifying finely-grained moments and infrequently-used strategic actions. However, of the remaining twenty-seven strategies from table 4.1, a further eight strategies (4, 9, 14, 15, 19, 22, 31 and 32) only recorded medium or higher in relation to *one* of the tests, and in some cases, a single item type, suggesting a very specific approach to those item types. This data is presented in table 4.10 below:

No.	Strategy	Test	Item type
4	Identifies grammatical or content-based parallel between items/options	IELTS	MC only
9	Marks/notes key prepositional phrase(s) in the question stem or options	IELTS	MC only
14	Marks/notes key adjective phrase in the text during careful reading	IELTS	MC and MI
15	Marks/notes key prepositional phrase in the text during careful reading	IELTS	MC only
19	Searches for key word/phrase (text)	IELTS	MC, SC, SuC
22	Identifies content-based parallel between paragraphs	IELTS	MC only
31	Checks/confirms response has met the parameters of the task	TOEFL	I-rp, I-it, R2L
32	Uses own topic knowledge to enhance understanding of text/questions	TOEFL	I only

Table 4.10. Strategies that recorded 'medium' or higher for only one of IELTS or TOEFL and relevant item type(s)

Of the eight strategies identified, six relate to IELTS and two to TOEFL. It is particularly noteworthy that multiple-choice items in IELTS multiple-choice (MC) items recorded very unique strategic approaches, accounting for all of the six unique strategies in addition to 'matching information' (MI), 'sentence completion' (SC) and 'summary completion' (SuC) items, which required test takers to focus specifically on word-level information to locate

relevant information. Strategies 4, 9 and 15 were largely unique to IELTS multiple-choice items. Two of these strategies were directly observable. Strategy 9 indicates that key propositional information is frequently contained within subordinate clauses in either the stem or options. Strategy 15 suggests that these items target detailed information contained within prepositional phrases in the text. Strategy 4 aligns with strategies 9 and 15, suggesting that participants are using complementary grammatical structures in the text and item in order to identify relevant information. IELTS multiple-choice items therefore target informationally dense parts of the text which must be parsed to successfully complete these items, and that this density is replicated in MC item design. However, the extent to which this item type targets higher or lower level processing will specifically be addressed in the next section. Targeting dense parts of the text is not a guarantee that the items target more complex, higher-level processing. Both 'multiple-choice' and 'matching information' items in IELTS specifically revealed participants' focusing on adjective phrases, suggesting that despite the intention of MI items to focus on main points of a paragraph, identification of the relevant key required a focus on descriptive information in addition to propositional information.

Despite the TOEFL being composed of mainly multiple-choice items, participants nonetheless behaved quite differently from multiple-choice items in IELTS, suggesting that MC items, despite being superficially similar, can be manipulated to target very specific strategic behaviour. Observable strategic behaviour provides valuable clues to underlying cognitive behaviour, but are not necessarily deterministic of specific levels of processing. Strategic behaviour can offer hypotheses for areas of difference between IELTS and TOEFL, which the data in table 4.10 provides.

For instance, strategies 22 and 32, which are partially predicated on verbal explanations to explicitly identify them (see section 4.2.1), suggest that IELTS MC items and TOEFL inferencing (I) items target high-level processing as the source of information for the observable actions is above the sentence level. Strategy 22 refers to observable instances of participants moving and comparing information between paragraphs. This behaviour is suggestive of linking propositional information. Verbal evidence is required to determine if this is the case, or whether the participant has linked word-level information across

paragraphs. Research question 3 specifically examines the cognitive processing that occurs in each item type and will explicitly address whether these item types target high-level processing.

Further similarities and differences emerged between Cohen and Upton's (2006) study and the present study. An interesting difference occurred in eliminating options as a strategy. Strategy 25 (Eliminates option(s) (contradicts key noun/verb phrase/nonsensical response) was used sporadically in this study, whereas the equivalent strategy (T28) is uniformly coded 'very high' in Cohen and Upton's study. In the present study, it was only 'very high' in relation to TOEFL 'inferencing' and 'reading to learn' items, suggesting that this strategy was consciously employed in relation to the more cognitively complex items. Option elimination was coded based primarily on video data, during which instances of test takers purposefully crossing out options was coded. These moments were then highlighted to the test taker, who was asked why they eliminated that option. Their response was then used to code this strategy as either 25 or 26. Their response was also analysed to determine what level of processing they had used in order to eliminate that option, which then contributed to the cognitive profile of that item. For Cohen and Upton, the strategy was coded based on a combination of video and verbal data. It is likely that participants' verbalisations were coded even when no observable elimination occurred. As the focus in their study was on the strategies rather than the cognitive processes underlying the elimination, this discrepancy between the studies is unsurprising.

It is clear form this discussion that the differences between Cohen and Upton (2006) are predicated on *observability* of strategic decision-making. The procedure for identifying strategies in the present study was primarily video data, and the purpose was to clearly identify strategic behaviour to use as stimuli for participants to talk about what they were thinking at explicit moments. For Cohen and Upton, strategy identification was the main purpose of their study. The different research objectives of the studies resulted in different outcomes.

4.2.6.1. Summary

Following Bachman and Palmer (1990, 1996, 2010) and Phakiti (2008), this study separated strategic competence from language ability (cognitive processing) for analytical purposes. Participants have command of the test management decisions they make and these may be observed or reported by the participants and researchers. A video-based approach to strategy identification resulted in very finely-grained analysis, allowing differences and similarities between IELTS and TOEFL to emerge. This study reported the positive identification of thirty-four strategic actions across used by the participants in IELTS and TOEFL, although the frequency with which each strategy was employed varied dramatically across the tests and across item types. Some strategies are used only sporadically, others persistently. A difference in research aims between the current study and Cohen and Upton drove differences in methodology which resulted in different outcomes in strategy identification. The strategies identified here represent the observable manifestations of test taker engagement with the tests. This aligns with what Phakiti calls 'strategic regulation' (2008). These observable actions regulate and moderate the cognitive engagement with the tests. As such, these actions constitute stimuli to present to test takers in order to acquire verbal evidence of their cognitive engagement with the tests.

The next section contains analysis of the processing core based on the verbalisations that emerged in the research. The best way to analyse this data was not by distinguishing between metacognitive and cognitive statements, which is ambiguous in a retrospective study. All verbalisations in a stimulated-recall interview are to some extent metacognitive, as the participant is being asking to recall what they were thinking at a specific moment. Thus the verbalisation cannot reflect precisely the thought processes at the moment that the test taker is recalling. The distinction between what they claim was their thought at that moment and what they think now about that moment is difficult to determine and an inherent weakness of stimulated-recall interviews. Instead, the focus is on the inferences that can be made about the participants' level of processing based on their verbalisations. The analysis focuses on higher versus lower level processing as identified in the processing core of the Khalifa and Weir (2009) model, and the constituent levels that make up these two categories. For the student model, the data is analysed commensurately with strategic data; by adding up the number of instances of each instance of processing for each item and then divide that total by the number of items of that type. Secondly, totals of higher and

lower level processing for each test will be determined. This will be transformed into a percentage. The relationship between any metacognitive activities and respective elements of the processing core will be explored.

4.3. Research question 2: Which cognitive reading processes do test takers use when they complete IELTS and TOEFL iBT reading sections?

Research question 1 was concerned with the identification of the strategies that participants used when they engaged with the two tests. These were the observable moments of engagement which served as the basis for subsequent interviews. Research question 2 is the first of three research questions which is concerned with the *cognitive* processes which participants used when they engaged with the two tests. This research question is explicitly concerned with the identification of cognitive processes, and addresses the question of which levels of cognitive processing form Khalifa and Weir's (2009) model can be inferred, based on participants' verbalisations.

Section 3.3.4.1 in the methodology chapter reported on the development of a coding schema. This section provides verbal evidence from participants to demonstrate how the coding schema was applied to individual verbalisations at each of the levels. The purpose of this section is therefore to demonstrate the utility of the model as a coding schema. The first part of this chapter provides verbal evidence for each of the levels of the model that can be used as the coding schema for comparability purposes, and verbal excerpts for each to demonstrate how the coding scheme was applied. For reasons of length, and because research question 2 is not directly concerned with comparability, only one example will be included for each of the levels of the coding schema. As stated in section 3.3.4.1, the coding schema includes two additional levels which are not specified in the original model. Each of these is included here, and the reasoning for the creation of two additional levels made explicit. This section concludes with a discussion on how changes to the model may influence future thinking into research and theoretical development into readers' cognitive processing.

4.3.1. Identification of cognitive processes used in the completion of IELTS and TOEFL

This section will demonstrate that from participant verbalisations we can infer the level of cognitive processing, based on whether the unit of reference is a single word, a phrase, a complete sentence, an understanding derived from adjacent sentences, an impression formed from engaging with an extended text, or in response to a level of understanding that has been established from multiple points of reference across multiple paragraphs culminating in a mental model. The level of understanding can be judged by the extent to which the participant rephrases parts of the text in their own words. Thus Khalifa and Weir's (2009) model may be applied to participant verbalisations via its constituent levels. This section will demonstrate that the model is analytically appropriate as it can account for the full range of verbalisations elicited in the SRI. Table 4.11 below is reproduced here from the methodology for quick reference:

	Level of processing	**Code**
Higher-level processes	**Creating an intertextual representation**: construct an organised representation across texts	**[P8]**
	Creating a text-level representation: construct an organised representation of a single text	**[P7]**
	Building a mental model: Integrating new information; enriching the proposition	**[P6]**
	Inferencing: At word/sentence/clause level	**[P5s/c]**
	Inferencing: At word level	**[P5w]**
Lower-level processes	**Establishing propositional meaning**: At sentence level	**[P4s]**
	Establishing propositional meaning: At clause level	**[P4c]**
	Syntactic parsing	**[P3]**
	Lexical Access	**[P2]**

Table 4.11. Final coding scheme for processing core of Khalifa & Weir's (2009) model of reading

The model contains nine levels. From the base to the top, the levels represent increasingly complex cognitive processing, based on the idea that cognitive processing becomes more

complex as informational density increases. Thus, the lexical basis for each level increases as one moves up the model. The lexical basis for each level is elaborated in section 3.3.4.1. The lowest category (word recognition [P1]) was removed from the model, as it was not considered theoretically interesting.

Two changes were also made to the model. Level four (establishing propositional meaning) occurs in the rubric at *clause* and *sentence* levels. Level five (inferencing) is also separated into two categories (inferencing at the word and clause/sentence level). 'Inferencing at the word level' refers to instances of participants integrating meaning across sentences at the word level, such as anaphoric and cataphoric referencing. Cataphoric referencing (referring to pronouns in previous sentences) is cited by Khalifa and Weir (2009: 51) as evidence of *inferencing* (a higher-level process) as it involves mentally linking information across two sentences which is not explicitly stated. Verbal evidence that they have mentally established inexplicit links between sentences can be coded 'inferencing' in Khalifa and Weir's model. Despite inferencing being a higher-level process in the model, this process only requires the linking of two lexical items in adjacent sentences. More complex inferencing involves mentally linking propositional ideas across sentences which are not explicitly linked in the text. This suggests that inferencing should be split into two categories; 'inferencing at the word level' and 'inferencing at the sentence level'. Coding verbalisations accordingly adds an additional dimension to test comparison to determine the *nature* of inferencing in the two tests. This difference is elaborated in participants' verbalisations in sections 4.3.1.5 and 4.3.1.6 below.

The remaining categories are consistent with Khalifa and Weir's model (lexical access; syntactic parsing; building a mental model and creating a text-level representation). The following sections detail how each of these codes outlined in table 4.11 was applied to verbal data. Each section defines the code and clarifies how it is applied to a verbalisation and provides an explicit example. The example is comprised of a transcribed verbalisation, the question it applies to, the relevant textual reference and observed behaviour that the transcription describes. Commentary on the example demonstrates how the code applies to the transcribed verbalisation.

4.3.1.1. Lexical access

Lexical access (P2) refers to instances in which participants select a response on the basis of identifying key words in item stems, options and finding equivalent or identical lexis in the text. This code was used in instances in which participants were observed underlining, circling or otherwise highlighting individual words in item stems, closely followed by similar actions undertaken in the text. Underlining a portion of the text is not sufficient evidence to claim that the participant has processed that portion. Observable actions such as underlining words served as the prompt in the SRI. Participant verbalisations then confirmed that *lexical access* underpinned this action. This code was also employed in more complex instances in which little direct action was observed.

Text reference: "…the high mountains of the Himalayas are only about 50 million years old. Lower mountains tend to be older, and are often the eroded **relics** of much higher mountain chains."

30. The word **relics** in the passage is closest in meaning to
- Resemblances
- Regions
- *Remains
- Restorations

Observed behaviour of participant 3 for item 30:
- Answers six seconds after answering item 29
- Participant selected option 3 for item 30 (correct) without returning to the text.
- Therefore, either used her existing knowledge of the text held in working memory, OR she knew the answer based on her 'lexical bank'.

Participant verbalisation:
"'Relics' means 'remains'. I didn't consider other options." [Participant 3, test 2, item 30]

For example, in item 30 in the TOEFL test, participant 3 is observed answering item 29, and then six seconds later, item 30. This action may be due to two reasons. Either she knew the answer to the vocabulary item (which may be completed without reference to the text), or she used a higher level process such as keeping extended textual information in her working memory. The verbalisation clarifies that it is the latter. She is immediately aware that the highlighted word ('relics') is synonymous with 'remains' (noun) in the context. Evidence that for this option, the participant only required 'lexical access' (P2).

4.3.1.2. Syntactic parsing

Syntactic parsing (P3) refers to verbalisations in which participants state that function words or sentence structure assisted them in getting the correct response; they cite no higher-level understanding of the part of the text they have engaged in. Use of this code was therefore prompted by similar observed behaviour as lexical access. In this instance, participants highlight lexical items other than key noun or verb phrases. Instead they focus on words on the basis of grammatical structure to assist them in relation to an item.

Text reference: "Causing participants in experiments to smile, for example, leads them to report more positive feelings and to **rate** cartoons (humorous drawings of people or situations)."

22. The word **rate** in the passage is closest in meaning to
- *judge
- reject
- draw
- Want

Observed behaviour of participant 5 for item 22:
- 00:42 Q9: Circles 'rate'
- 00:43 Circles 'rate' in text (paragraph 4, line 3)
- 00:44 Q9: Circles 'closest'
- 00:50 Underlines 'humorous drawings of people or situations'
- 00:54 Q9: Considers options
- 01:00 Circles 'report' (paragraph 4, line 3)

Participant verbalisation:
"…because there's 'and' here, so I was asked to answer this; 'rate', and here, 'report' is in the same position, so it should be the same function, so here it's 'report', then it says it's 'report the feeling', so here is should be 'rate cartoons'… 'report' is like 'describe', like 'describe feelings'" [participant 5, test 1, item 22].

For example, in relation to item 22, participant 5 is observed circling the word 'rate' in both the stem and text (lexical access). Then she circles 'closest' and a noun phrase, consistent with identifying the purpose of the question. She then identifies 'report', which seemingly bears little significance to completing the item. When asked to explain this action, she explained that this word is fulfilling a similar function as 'rate'. The sentence structure is the same, indicating that the structure provided contextual cues as to the purpose of the word 'rate' in the text. This was therefore coded 'syntactic parsing'.

4.3.1.3. Establishing propositional meaning at the clause level

Establishing propositional meaning at the clause level (P4c) refers to participant responses in which they clearly respond on the basis of understanding clause-level information (contained within a single clause – a subject, a verb and an object), which may be a complete sentence or part of a longer, multi-clause sentence. This level of processing requires the observation of participants reading part of the text carefully, then, when prompted, being able to paraphrase or explain the propositional content of a single clause.

Text reference: "In classic research Paul Ekman took **photographs** of **people** exhibiting the emotions of anger disgust, fear, happiness, and sadness. He then asked **people** around the world to indicate what **emotions** were being depicted in **them**."

17. The word **them** in the passage refers to
- emotions
- people
- *photographs
- cultures

Observed behaviour of participant 2, item 17:
Participant spends more than one minute focusing on above portion of the text.

P**articipant verbalisation:**
"I see in the sentence, 'in classic research, Paul Ekman took photographs of people', which means this phrase [photographs] is the main word in this phrase. So that is reassuring that I might make the right choice. They are not talking about people, they are talking about [photographs]" [participant 2, test 2, item 17].

In the above example, the participant was observed engaging with the text reference for more than one minute without making any observable action. When prompted to explain what they were thinking at that moment, the participant was able to explicitly state which sentence they were engaging with, and specifically which part had prompted them to make the connection to the item stem. She cites the clause 'took photographs of people' and states that 'photograph' is the main noun. As the cited information is contained within a single clause, this is coded 'propositional meaning at the clause level'.

4.3.1.4. Establishing propositional meaning at the sentence level

Establishing propositional meaning at the sentence level (P4s) is similar to the previous code, although it extends the frame of textual reference above the clause level. Instances of this behaviour are regarded as more complex than the verbalisations linked to single clauses, as the participant is required to process more information. Specifically, the code is used when the participant has clearly had to parse more than one verb, subject and/or object.

Text reference: "Hills and mountains are often regarded as the epitome of permanence, successfully resisting the destructive forces of nature, but in fact they tend to be relatively short-lived in geological terms."

35. Which of the sentences below best expresses the essential information in the highlighted sentence in the passage?

Incorrect choices change the meaning in important ways or leave out essential information.

- When they are relatively young, hills and mountains successfully resist the destructive forces of nature.
- *Although they seem permanent, hills and mountains exist for a relatively short period of geological time.
- Hills and mountains successfully resist the destructive forces of nature, but only for a short time.
- Hills and mountains resist the destructive forces of nature better than other types of landforms.

Observed behaviour of participant 6, item 35:
01:56 Q35: Begins reading stem and options
02:12 Q35: Underlines 'best expresses'
02:15 Moves to text (paragraph 2). Reads highlighted sentence in context
03:25 Q35: Selects option 2

Participant verbalisation:
"I think the option 2 is most suitable, is most similar with the meaning of the sentence, so I choose option 2. The sentence says 'hills and mountains are often regarded as permanent, but…'. It's [this] 'but', in fact they tend to be relatively short-lived', so this, option 2 is in the same format, 'although', they are permanent, 'but' 'they exist for a relatively short geological time'".
[Participant 6, test 1, item 35]

In relation to item 35, participant 6 uses lexical access to identify the relevant part of the text for the item. He then shows evidence of reading this part of the text in depth. He selects option 2. In explaining his actions, he identifies the key word 'but' and 'although' as

linking parts of the sentence together, and displays evidence of linking both clauses in the key to both parts of the sentence in the textual reference. The sentence contains multiple verb and noun phrases to parse accurately in order to successfully answer this item.

4.3.1.5. Inferencing at the word level

Inferencing at the word level (P5w) is the first of the levels from the Khalifa and Weir model that is labelled a 'higher-level' process. This is because it refers to instances of participants establishing meaning *across* sentences. Inferencing was divided into two individual codes, the first of which relates to instances of participant behaviour where they establish meaning across sentences at the *word* level. This refers to anaphoric and cataphoric referencing, that is, linking a noun to its referent pronoun. This is the lowest level of inferential reasoning. It is cited as 'inference' as participants must determine which noun phrase links to relevant pronouns, which requires propositional understanding in more than one sentence.

Text reference: "Under very cold conditions, rocks can be shattered by ice and frost. Glaciers may form in permanently cold areas, and these slowly moving masses of ice cut out valleys, carrying with them huge quantities of eroded rock debris."

34. The word **them** in the passage refers to
    - cold areas
    - *masses of ice
    - valleys
    - rock debris

Observed behaviour of participant 6, item 34:
00:10 Begins reading text (paragraph 6) carefully
01:00 Q34: Begins reading stem and options
01:06 Q34: Selects option 2

Participant verbalisation:
"'them' here is connected to the first part of the sentence, so it mentions that glaciers carry the masses of ice and then I think 'them' is continuing to talk about the masses of ice. I don't know how glaciers can carry valleys, 'rock debris' is… because it's carrying with rock debris, it's not very correct, so I just chose 'masses of ice'. I just put 'masses of ice' into the sentence, replacing [them] to read it again."

For item 34, participant 6 did not display much observable behaviour, although answered the item rapidly (less than one minute). When asked to explain his thought process that led to this action, he stated that he was able to select the correct option by mentally placing it in the position of the pronoun and thus formed propositional meaning across the two

sentences. Other options did not allow him to establish propositional meaning. He correctly recognises that the pronoun refers to an object that 'carries' another and that this cannot therefore apply to the valleys. *Inferencing* in Khalifa and Weir's model is identified as a single category, although in the exposition offered (2009: 50-51), is described as operating on three levels; text, clause/sentence and word. Text-level inferencing is described as "imposing coherence" on a text by schematic knowledge of a reader is added to the explicitly-stated information in a text to create an additional level of meaning that is not explicitly-stated. Clause or sentence level inferential processing specifically cites "anaphor resolution", in which an anaphor (such as a pronoun) refers to the antecedent in the preceding clause or sentence. The anaphor on its own makes little sense conceptually so must be understood in terms of the preceding information. Anaphors and antecedents may be in the same sentence or several sentences apart. The further an anaphor from its antecedent, the more difficult the text is to parse. Word-level inferencing occurs when a word with multiple meanings (a homograph) needs to be understood in its context; this context will supply the information to the reader to select the appropriate meaning. Anaphoric referencing (clause/sentence level inference) is encapsulated in the next section.

4.3.1.6. Inferencing at the sentence/clause level

Inferencing at the sentence/clause level (P5s/c) refers to instances in which participants establish meaning in two or more adjacent sentences (each of which may contain one or more clauses) and then verbalising a proposition which is *not explicitly stated in the text* on the basis of this understanding. This is a much more complex cognitive operation, as participants are expected to form an additional mental proposition on the basis of two explicitly state ones.

Note that the demands on participant verbalisations increase substantially from this level. In order for higher level codes to be used, participants have to provide extended explanation to satisfactorily demonstrate they have processed at this level. Processing at higher levels is also qualitatively different. Forming an unstated proposition is not reading per se, but is a cognitive process made on the basis of accumulated information acquired from reading.

Extract from paragraph 2: As a general rule, the higher a mountain is, the more recently it was formed; for example, the high mountains of the Himalayas are only about 50 million years old. Lower mountains tend to be older, and are often the eroded relics of much higher mountain chains. About 400 million years ago, when the present-day continents of North America and Europe were joined, the Caledonian mountain chain was the same size as the modern Himalayas. Today, however, the relics of the Caledonian orogeny (mountain-building period) exist as the comparatively low mountains of Greenland, the northern Appalachians in the United States, the Scottish Highlands, and the Norwegian coastal plateau.

Q29. Which of the following can be inferred from paragraph 2 about the mountains of the Himalayas?

- Their current height is not an indication of their age.
- *At present, they are much higher than the mountains of the Caledonian range.
- They were a uniform height about 400 million years ago.
- They are not as high as the Caledonian mountains were 400 million years ago.

Participant verbalisation:

"I think this sentence is correct from the information that I have found…as a general rule, the higher a mountain is, the more recently it was formed. So, because the age of the Himalayas is younger, so it's higher than the Caledonian range" [Participant 6, test 1, item 29].

In his verbalisations for item 29, participant 6 repeats the first sentence of paragraph 2 verbatim, then states a general rule that higher mountains are younger, summarising sentence 2 in paragraph 2. He then states that the Himalayas are higher than the Caledonian mountain range; therefore, the Himalayas are younger. The first two statements are direct retellings of the first half of the paragraph. The third statement is inferred on the basis of the first two. The participant established propositional meaning in two sentences, then established a third proposition. This code was used for instances in which participants made statements that were not directly given in the text. They need to be able to verbalise how they arrived at this statement.

4.3.1.7. Building a mental model

Building a mental model (P6) refers to verbalisations that summarise portions of the text and demonstrate that the participant was able to differentiate main from supporting details. Building a mental model differs from inferencing at the sentence level in two crucial ways. First, when building a mental model, participants are not required to discern an unstated propositional statement from what they have read. Secondly, the text load

required to provide evidence of a mental model is larger. In inferring at the sentence level, participants are required to establish propositional meaning of two sentences. To provide evidence of building a mental model, participants must be able to summarise a larger portion of text; three or more sentences in which the participant is able to state the main purpose of the part of the text, and how additional sentences cohere to provide additional or supporting information. An example is given below.

Paragraph 5: The weather, in its many forms, is the main agent of erosion. Rain washes away loose soil and penetrates cracks in the rocks. Carbon dioxide in the air reacts with the rainwater, forming a weak acid (carbonic acid) that may chemically attack the rocks. The rain seeps underground and the water may reappear later as springs. These springs are the sources of streams and rivers, which cut through the rocks and carry away debris from the mountains to the lowlands.

32. Why does the author mention Carbon dioxide in the passage?

- *To explain the origin of a chemical that can erode rocks
- To contrast carbon dioxide with carbonic acid
- To give an example of how rainwater penetrates soil
- To argue for the desirability of preventing erosion

Participant verbalisation:
"'Carbon dioxide' is a kind of chemical, and it can erode rocks, but does the article talk about the origin of this chemical? I don't recall the article talking about that… I read the sentence again when it talks about carbon dioxide, and certainly not comparing the two of them, because carbonic acid is simply the outcome after the interaction with the chemical and the rocks. What does it mean to say that rainwater penetrates soil? Because the previous sentence talks about 'rain washes soil and penetrates cracks', so this sentence still a continuation of that… thinking about whether it's a continuation of that topic or it's an initiation for a new idea, so I had to read the sentence again, and then I decided it's a continuation of the idea, so I sort of confirmed with myself the choice" [Participant 3, test 2, item 32]

Item 32 in the TOEFL test is a rhetorical purpose item. Participants are required to understand why a specific word or phrase is mentioned in the text. In her explanation of why she selected option 1, participant 3 explains why she did not select option 2-4. In doing so, she demonstrates that she understands that the sentence containing 'carbon dioxide' is continuing the idea contained within a previous sentence, explaining how rocks are eroded. This requires the establishment of propositional meaning across three sentences; the weather as the main agent of erosion, rain penetrating rocks, and carbon dioxide forming, which attacks the rocks. Complex ideas included in the distractors require the participant to eliminate these options and lead her to understanding the purpose of the paragraph and why carbon dioxide is mentioned.

4.3.1.8. Creating a text-level representation

The final code, creating a text-level representation (P7) is more difficult to utilise. It requires participants to verbalise understanding of several main points in the text and the purpose of the text itself. *Text-level representation is unlikely to reveal itself in single verbalisations.* This code was used when participants have verbalised several instances of building a mental model and then demonstrated text-level understanding when they were able to clearly state what the purpose of the text is and can locate main arguments throughout the text without significant conscious effort. It was not anticipated that this code would be utilised much; it was used to test claims made by test developers that specific item types require participants to form a text-level representation, or whether participants were able to respond correctly using lower levels of processing.

Directions:
38. Three of the answer choices below are used in the passage to illustrate constructive processes, and two are used to illustrate destructive processes. Complete the table by matching appropriate answer choices to the processes they are used to illustrate. This question is worth 3 points.

Constructive Processes	**Destructive processes**
1.	**1.**
2.	**2.**
3.	

Answer Choices

1. Collision of Earth's crustal plates 2. Separation of continents 3. Wind-driven sand 4. Formation of grass roots in soil 5. Earthquakes 6. Volcanic activity 7. Weather processes

Participant verbalisation:
"I remember when I was reading the article at the very beginning, I noted down a few key words, ***and in one paragraph in particular****, it talks about three different constructive processes, and I have impressions of them being crustal plates, earthquakes and volcanic activity, so that's* ***why I very quickly picked out 1, 5 and 6.*** *I did go back to the text when I was deciding on number 3. For number 7 I decided that very quickly because* ***I remember one of the main destructive forces mentioned was weather processes****. I eliminated number 4 because the 'formation of grass roots' is actually part of the forces that protect rocks from erosion. But then I was thinking about separation of continents and thinking whether in the texts it talks about separation of continents within the discussion about destructive processes. This phrase, to remind myself as cue that from this moment onwards,* ***the author begins to talk about destructive forces****."* [Participant 3, test 2, item 38].

Item 38 is a 'reading-to-learn' item. Test takers are required to organise key points in the text. In this example, test takers are presented with a text on geology and are asked to identify three constructive processes and two destructive processes which are cited in the text. The participant was able to respond to this item without returning to the text. The verbalisation confirmed that she had formed a mental representation of the text which she was able to relay in interview. She is able to cite one paragraph, the purpose of which was to identify key constructive processes. She recognises option 4 as a supporting detail in discussing erosion. She is also able to pinpoint the moment within the text that the author changes focus. This verbalisation demonstrates that she has combined information from multiple points in the text.

4.3.1.9. Creating inter-textual representation

No instances of the final code (P8) were recorded for either test. This is not unexpected, as neither test contains a task which asks participants to combine information from more than one text. All items in both tests relate to a single text only. A focus on inter-textual representation would require two (or more) texts on a single topic so that participants can formulate links across them. As the IELTS test contains only three texts, this would mean that the range of texts presented to test takers would be limited in genre, coupled with the difficulty of designing and specifying item types that require participants to form text-level representations of more than one text, and then demonstrate that they can compare information across texts in some way. This would likely require some form of extended writing activity to demonstrate command of both texts. The original 1989 version of IELTS featured an integrated approach between the IELTS reading and writing modules. This link was removed in the 1995 IELTS revision project because it increased the potential for confusing assessment of writing ability and assessment of reading ability (Charge and Taylor, 1997). Monitoring of candidates' writing performance suggested that the extent to which they exploited the reading input also varied considerably, and that the cultural background of candidates could be a factor in this.

IELTS Revisions in the past thirty years have been towards less integration of individual sections. A previous version of IELTS, the English Language Testing System (ELTS) contained six subject-specific reading tests that could be undertaken. In 1988, this was reduced to three (Alderson and Clapham, 1992: 1). In 1995, a further revision occurred with three subject-specific tests being replaced with one academic reading test. Charge and Taylor (1997) identify three reasons for this change; first, individual institutions found placing students in the correct module to be arbitrary, the divisions between the modules to be ambiguous, second, there was a conflict between placing students in a particular course based on their background or their intended course of study, and third, 75% of candidates were opting for Module C, the Business Studies and Social Science module. (Charge and Taylor, 1997). Clapham (1996) suggested that one academic reading paper "would not discriminate for or against candidates of any discipline area." It is therefore extremely unlikely that IELTS will move towards testing inter-textual representation in the near future. Despite the clear advantages of moving away from items types that involve multiple sources in both test administration and scoring, this has resulted in a truncated definition of the construct of reading for academic purposes that omits one of the most fundamental aspects of higher education; integrating multiple sources of information and forming an argument on the basis of this.

In contrast, the revised TOEFL-iBT was underpinned by the development of a conceptual framework that takes into account models of communicative competence (Jamieson et al, 2000) and is promoted on the basis of "integrated" tasks in which productive activities relate to a combination of listening/reading inputs, although ETS reports scores for each of the four skills. As the scope of this PhD is limited to a single skill (reading), the integrated-skills items are not being considered for analysis. IELTS does not contain integrated-skills items so there is no scope for comparative content analysis of these items. Integrated skills tasks in the TOEFL test re based on integrating reading, listening and writing, and therefore do not contain multiple writing sources form which to derive information. Therefore, the criticism that IELTS has a limited definition of reading for academic purposes also holds true for the TOEFL iBT.

4.3.2. Reflections on the application of the coding schema to the data and theoretical contribution to the construct of L2 reading

The use of a coding scheme developed from Khalifa and Weir's (2009) cognitive processing core invited critical reflection on the applicability of these codes to participant verbalisations based on the working definitions of each of the processing levels used in this study. The bottom four layers of Khalifa and Weir's model are presented as 'lower-level processing' and the top five layers as 'higher-level processing'. As one moves up the core, the cognitive demands on the reader increase. Lower-level processing initially requires decoding at orthographic and phonological (sub-word) levels. Orthographic and phonological knowledge contribute to overall word identification (lexical access). This is known as 'word form' (Nation, 2000). Grammatical and syntactic competencies allow readers to form meaning across words to derive meaning from lexical strings (syntactic parsing). Sentence and clause processing requires morphosyntactic structural knowledge to understand how each word contributes to propositional meaning (Anderson and Nagy, 1991). Participant verbalisations clearly demonstrated that the linguistic basis of their decision-making could be demarcated according to the four levels which constitute lower-level processing in Khalifa and Weir's model.

Higher-level processing in this model comprises inferencing, building a mental model, creating a text-level representation and creating intertextual representation. Inferencing at the clause or sentence level refers to instances in which participants establish meaning in two or more adjacent or non-adjacent sentences (each of which may contain one or more clauses) and then verbalising a proposition which is formed from an understanding of those two clauses or sentences which is not explicitly stated in the text. Building a mental model refers to the process by which the reader forms a mental summary of portions of the text. Creating a text-level representation requires readers to understand several main points in the text and the purpose of the text itself. Intertextual representation in this model requires readers to form text-level representation across two (or more) texts, and then elaborating on how their understanding of one has informed the other. This suggests that both tests cover the same range of cognitive processes, although this data does not demonstrate to what extent prominence is given to the different levels. This will be explored in the next

section. Furthermore, the latter, crucial component of reading for academic purposes is not included in the construct definition for either IELTS or TOEFL.

From the investigation into test takers' cognitive processes in this thesis, it is clear that the distinction between lower and higher-level processing could benefit from further elaboration and investigation. The boundary between lower and higher-level processing in Khalifa and Weir (2009) lies at the sentence boundary. Within sentences, knowledge of grammar, syntax and vocabulary is sufficient to discern the writer's propositional meaning. However, above the sentence level, the point at which participants are required to link ideas across sentences, this knowledge is insufficient to comprehend the writer's argument. This is the reason that inferential reasoning is cited as a higher-level process in the model, as it implies formulating an idea based on two or more propositions which is not explicitly stated in the text. This is the definition of inferential reasoning offered by Jamieson et al (2008: 244) for the TOEFL test:

> "…comprehend an argument or idea that is strongly implied but not explicitly stated in the text, identify the nature of the link between specific features of exposition and the author's rhetorical purpose…"

However, the definition of 'inferencing' offered by Khalifa and Weir includes *cataphoric referencing* (linking a noun to its pronoun in a previous clause or sentence) as it involves mentally linking two lexical items in adjacent (or non-adjacent) clauses or sentences (2009: 51). This suggests that the transition between higher and lower-level processing may occur at the *sub-sentence level as well as above the sentence level*, as sentences may contain more than one clause (e.g. relative or subordinate clauses), with pronominal referencing between clauses. Alternatively, two simple short sentences may contain pronominal referencing linking them together. Consider the textual reference from section 4.3.1.4 above:

> *"**Hills and mountains** are often regarded as the epitome of permanence, successfully resisting the destructive forces of nature, but in fact **they** tend to be relatively short-lived in geological terms."*

In this case, the pronoun 'they' in the third line refers to 'hills and mountains' in the first line. This pronominal reference is separated by a subordinate clause ('successfully…') meaning that the cognitive processing required to process this single sentence may exceed that of multiple simple sentences containing similar information. The further the distance between the noun and its pronoun, the more complex the cognitive processing required to make the cognitive link.

This issue is mitigated by the creation of two additional coding levels in order to recognise and address these concerns. The addition of two codes at the boundary between lower and higher level processing acknowledges the nuance found at this boundary, by identifying instances of processing which occur at the *clause* level (identifying propositional meaning), the sentence level (in which propositional meaning is gained in a sentence which includes multiple clauses), inferencing at the word level (to account for pronominal referencing occurring within or across sentence boundaries) and inferencing at the clause/sentence level, to account for instances of test takers forming a proposition which is not explicitly stated in the text.

For the purposes of clarity and consistency with existing literature, the boundary between higher and lower-level processing in the current study remains at the sentence boundary (P4s/c are lower and P5w/s/c are higher-level processing). However, this does not resolve the issue of whether propositional understanding of complex sentences (coded P4s) can be regarded as a higher level of processing than a basic form of inferential reasoning such as linking pronouns across sentences (coded P5w). Arguably, higher-level processing should be regarded as having commenced at the moment when participants are required to connect ideas *across clauses*. As demonstrated above, individual sentences can be semantically complex and contain multiple interconnected ideas which the reader must parse to gain full understanding of the writer's intended meaning. The clause boundary acts as a much clearer demarcation of the moment at which grammatical and morphosyntactic structural knowledge is no longer sufficient to fully account for propositional understanding.

The use of Khalifa and Weir's (2009) model to analyse test taker behaviour allows researchers to specify participant behaviour with a high degree of accuracy. Khalifa and

Weir's model effectively amounts to a hypothesis of how participants form meaning from the text with which they are engaging. Higher and lower levels of the model are fundamentally different. Principles of grammar and syntax apply to the first four levels, with the higher levels being fundamentally organisational in nature, with the idea or concept as the principle unit of analysis. The higher levels of the model represent the mental formation of ideas based on parsing text, with the lower levels as the linguistic components and subject to the metalinguistic awareness of the participant. The lower levels of the Khalifa and Weir model are highly automatized as they imply linking words to the participant's lexical repository. Grabe (2009) argues that lower-level processes can become automatized as the participant becomes more proficient, as their innate language capacity is able to handle the full range of grammatical constructions, syntactic patterns and lexical items presented to them in any given text input. The higher levels of processing less so because they involve the formulation of ideas and meanings across sentences. These ideas are accessible to participants and the recitation of those ideas can be used to code verbalisations at specific levels of the Khalifa and Weir model. This requires rich and detailed stimuli which the methodology has provided. The verbalisations provided evidence for the hierarchical nature of the Khalifa and Weir model with the majority of codings in the low level processing categories.

This section has addressed research question 2, which directly sought to identify the cognitive processes that could be identified based on the adopted methodology. The findings revealed that the methodology was actually appropriate for identifying a greater number of levels than were initially proposed in Khalifa and Weir's (2009) model. Two additional levels were defined and identified, indicating that the methodology was successful in identifying finely-grained differences in levels of processing that participants employed to answer individual items. An example verbalisation was provided for each of the levels of the model. A further finding was that both tests had a truncated construct definition that does not encompass the highest level of Khalifa and Weir's model (creating inter-textual representation). This is an important aspect of reading for academic purposes, which is missing from both IELTS and TOEFL. The next section directly addresses research question 3 which seeks to explore the conception of the construct by the both sets of test developers more closely.

4.4. Research question 3: Do these processes reveal differences in IELTS and TOEFL iBT test developers' understanding of the construct of interest?

Section 4.4 directly addresses research question 3. This section is explicitly concerned with the comparability of the cognitive processing in IELTS and TOEFL. Research question 2 identified the cognitive processes that could be identified from participants' verbalisations. This section presents the outcomes of participant engagement with the tests in terms of their cognitive processing. The previous section detailed the processes that could be inferred from participants' observable engagement with the test instruments without reference to whether participants were successful or unsuccessful. This section therefore commences with the outcomes of the six participants in terms of their success. Responses to individual items are presented for all participants for both the IELTS and TOEFL tests. This makes clear the necessity of having two complete verbal accounts for each test, to increase the likelihood of recording a correct response for each item.

Secondly, this section presents the detailed evidence for the levels of cognitive processing used by the test takers. Research question 3 relates to the student model, which describes the construct of interest. Research question 3 is aimed at discerning the nature of the construct in both the IELTS and TOEFL tests in terms of the levels of Khalifa and Weir's (2009) processing core. This section presents the respective frequencies of each of Khalifa and Weir's processing levels for IELTS and TOEFL, in order to compare directly the composition of each test. This section details the outcomes of the SRI at the *test level* in order to directly compare the cognitive processes that were encoded in both tests and make comparative claims. Similarities and differences will be identified, and the findings discussed in relation to the literature relating to cognitive processing for both tests.

4.4.1. Research participants' task completion success in IELTS and TOEFL

This section provides the test results of the participants in the research. The IELTS test contains forty items, each of which was completed by two test takers, meaning there are

eighty individual item responses collected in the research for the IELTS test. Of these, 23 responses were incorrect, and 57 were correct (71.25 per cent). The TOEFL test contains 38 items, although as each section contains a partial credit, reading-to-learn item, the total number of options to be selected is 46 for the full test, 92 in total for the two administrations of the TOEFL. Of these, 14 are incorrect and 78 correct (84.78 per cent). The participants were generally more likely to answer correctly on the TOEFL test than the IELTS test, despite having significantly less experience with the TOEFL. This is evidence that their lack of knowledge related to the TOEFL test did not prevent participants from providing meaningful responses to the TOEFL test and subsequent verbalisations.

The low-stakes nature of the research meant that the proportion of correct responses was lower than expected given the proficiency of the participants. No correct record was obtained for four out of forty items (8, 10, 12 and 19) in the IELTS test. Likewise, no correct record was obtained for one item (24) in the TOEFL test. Nonetheless, verbalisations relating to incorrect responses were retained for analysis as they were considered meaningful to address the issue of whether participants had answered incorrectly because they did not utilise the correct level of processing as the requirements of the task were beyond them, or whether some aspect of task design influenced their test management strategies resulting in them relying on an inappropriate level of processing. Instances in which one participant responded correctly while the other answered incorrectly were particularly illuminating to address this question. From the perspective of RE, this information can be extremely valuable to test design or re-design. The complete record of participant responses to the IELTS and TOEFL tests are presented in tables 4.12 and 4.13 below:

Test	Text title	Question	Participant Responses				
			Participant 1	✓ / X	**Participant 4**	✓ / X	**Correct Responses**
IELTS	**Striking Back at Lightning with Lasers**	1	D	✓	C	X	**D**
		2	A	✓	A	✓	**A**
		3	A	✓	A	✓	**A**
		4	Power companies	✓	Power companies	✓	**Power**
		5	safely	✓	safely	✓	**companies**
		6	s---	X	size	✓	**safely**
		7	B	✓	B	✓	**size**
		8	D	X	D	X	**B**
		9	H	X	G	✓	**C**
		10	A	X	C	X	**G**
		11	Not given	X	No	✓	**D**
		12	No	X	Not given	X	**NO**
		13	Yes	X	Not given	✓	**YES**
							NOT GIVEN
	The Nature of Genius	**Question**	**Participant 2**	✓ / X	**Participant 5**	✓ / X	**Correct Responses**
		14	B	✓	A	X	**B**
		15	E	X	B	✓	**C**
		16	G	X	C	✓	**F**
		17	H	✓	G	X	**H**
		18	F	✓	H	✓	**J**
		19	False	X	False	X	**TRUE**
		20	True	✓	True	✓	**TRUE**
		21	False	✓	False	✓	**FALSE**
		22	False	X	True	✓	**TRUE**
		23	Not given	X	True	✓	**TRUE**
		24	False	X	Not given	✓	**NOT GIVEN**
		25	True	✓	True	✓	**TRUE**
		26	Not given	✓	Not given	✓	**NOT GIVEN**
	How Does the Biological Clock Tick?	**Question**	**Participant 3**	✓ / X	**Participant 6**	✓ / X	**Correct Responses**
		27	iii	X	ix	✓	**ix**
		28	ii	✓	x	X	**ii**
		29	vii	✓	vii	✓	**vii**
		30	i	✓	i	✓	**i**
		31	viii	✓	viii	✓	**viii**
		32	iv	✓	iv	✓	**iv**
		33	physical chemistry	✓	physical chemistry	✓	**physical chemistry**
		34	thermodynamics	✓	thermodynamics	✓	**thermodynamics**
		35	adapt	✓	adapt	✓	**adapt**
		36	Immortality	✓	Immortality	✓	**Immortality**
		37	No	✓	Yes	X	**NO**
		38	Yes	✓	Yes	✓	**YES**
		39	Not given	✓	Not given	✓	**NOT GIVEN**
		40	Yes	✓	Yes	✓	**YES**

Table 4.12. Participant responses (IELTS)

Test	Text title	Question	Participant Responses				
			Participant 1	**✓ / X**	**Participant 4**	**✓ / X**	**Correct Responses**
TOEFL	**Nineteenth-century politics in the Unites States**	1	Option 2	✓	Option 2	✓	Option 2
		2	Option 3	✓	Option 3	✓	Option 3
		3	Option 2	✓	Option 2	✓	Option 2
		4	Option 1	✓	Option 1	✓	Option 1
		5	Option 3	✓	Option 3	✓	Option 3
		6	Option 2	✓	Option 2	✓	Option 2
		7	Option 3	✓	Option 3	✓	Option 3
		8	Option 4	✓	Option 4	**X**	Option 4
		9	Option 4	✓	Option 1	**X**	Option 4
		10	Option 2	**X**	Option 4	✓	Option 2
		11	Option 3	✓	Option 4	✓	Option 4
		12	Option 1	✓ (2/2)	Option 1	✓ (2/2)	Option 1
		13	Options 1, 5 and 6		Options 1, 5 and 6		Options 1, 5 and 6
	The Expression of Emotions	**Question**	**Participant 2**	**✓ / X**	**Participant 5**	**✓ / X**	**Correct Responses**
		14	Option 2	✓	Option 2	✓	Option 2
		15	Option 3	✓	Option 3	✓	Option 3
		16	Option 2	✓	Option 2	✓	Option 2
		17	Option 3	✓	Option 3	✓	Option 3
		18	Option 3	✓	Option 3	✓	Option 3
		19	Option 3	✓	Option 3	✓	Option 3
		20	Option 1	✓	Option 1	**X**	Option 1
		21	Option 1	✓	Option 2	**X**	Option 1
		22	Option 1	✓	Option 3	**X**	Option 1
		23	Option 4	**X**	Option 1	**X**	Option 4
		24	Option 1	**X**	Option 1	✓	Option 4
		25	Option 4	✓ (1/2)	Option 3	✓ (2/2)	Option 3
		26	Options 2, 5 and 6		Options 2, 4, and 6		Options 2, 4, and 6
	Geology and Landscape	**Question**	**Participant 3**	**✓ / X**	**Participant 6**	**✓ / X**	**Correct Responses**
		27	Option 4	✓	Option 4	✓	Option 4
		28	Option 2	✓	Option 2	✓	Option 2
		29	Option 2	✓	Option 2	✓	Option 2
		30	Option 3	✓	Option 1	**X**	Option 3
		31	Option 3	**X**	Option 3	✓	Option 3
		32	Option 3	✓	Option 1	✓	Option 1
		33	Option 2	✓	Option 2	✓	Option 2
		34	Option 2	✓	Option 2	✓	Option 2
		35	Option 2	**X**	Option 2	✓	Option 2
		36	Option 1	✓	Option 4	✓	Option 4
		37	Option 1	✓ (3/3)	Option 1	✓ (2/3)	Option 1
		38	Options 1, 5, 6; 3, 7		Options 1, 5, 6; 3, 4		Options 1, 5, 6; 3, 7

Table 4.13. Participant responses (TOEFL iBT)

Each table includes the title of each text and the numbers of the questions presented alongside that text. The tables include the selections of the participants, whether they responded correctly or incorrectly to each item, and the correct response for each item. Items that require more than one selection include all the relevant keys, and the number of correct responses by each participant for those items is given in brackets, as is the maximum score that can be awarded for those items.

Participants were recruited on the basis of their proficiency and were given more time to complete all items than in a live administration for two reasons. First, in an attempt to secure at least one correct response for all items, and to ensure that participants were given sufficient time on each item so that they would have something to contribute in the subsequent interviews. However, particularly in IELTS, participant scores were lower than expected. This was attributed to the low-stakes nature of the research. Participants conducted the tests in different sessions, ruling out fatigue as a possibility. Tobia and Everson (2009) suggest that in low-stakes research, participants may be less likely to be successful since they are unlikely to take extra actions or effort to improve their work or to use more advanced metacognitive strategies in the event that they are uncertain about a response. Monitoring accuracy (performance appraisal) will influence affective schemata and therefore cognitive processing. This is a weakness that the study attempted to overcome by amending the methodology so as not to overload participants, but this remains a significant drawback of the current study. The next section will outline the overall cognitive picture that emerged of both the IELTS and TOEFL tests.

4.4.2. Test takers' cognitive processing in completing IELTS and TOEFL

Table 4.14 below contains the identified cognitive processes from table 4.11 associated with both the IELTS and TOEFL tests. This table is the summary of coded verbalisations for all participants and items in both tests:

Higher/ lower	Cognitive Processing	IELTS	%	Higher/ lower	TOEFL iBT	%	Higher/ lower
Lower	P2	**93**	39.74		**108**	34.84	
	P3	**28**	11.97		**31**	10.00	82.58

	P4c	**40**	17.09	88.03	**66**	21.29	
	P4s	**45**	19.23		**51**	16.45	
Higher	P5w	**2**	0.85		**7**	2.26	
	P5s/c	**4**	1.71		**10**	3.23	
	P6	**20**	8.55		**33**	10.65	
	P7	**2**	0.85		**4**	1.29	
	P8	**0**	0	11.97	**0**	0	17.42
	Total	**234**	**100**	**100**	**310**	**100**	**100**

Table 4.14. Identified cognitive processes for IELTS and TOEFL

Five main claims are apparent from this data. First, the findings confirm that each of the levels of cognitive processing in Khalifa and Weir's model have been identified in both IELTS and TOEFL, with the exception of 'forming inter-textual representation' (P8), as outlined in response to research question 2. Second, the majority of identified cognitive processes were lower level processes for both tests. Third, the TOEFL test recorded a greater number of instances of each level of processing than the IELTS test. Fourth, the pattern of recorded cognitive processes across the two tests is consistent, which suggests that the codes have been consistently applied across both tests, and that they have a very similar conception of the construct of reading for academic purposes. Finally, high level processing in both tests lies primarily in 'forming a mental model' (P6) rather than inferencing or creating a text-level representation.

Eight of the nine levels in the coding schema have been identified for both IELTS and TOEFL. As clarified in section 4.3, no items in IELTS or TOEFL target the highest level of processing in Khalifa and Weir's model (P8), indicating a truncated conception of the construct of reading for academic purposes. This issue was discussed in the previous section. A second finding is that the majority of processes identified are lower level processes. For IELTS, 88.03 per cent of the processes identified were lower level. The figure for TOEFL was 82.59 per cent. There could be two explanations for this. First, that the tests overwhelmingly target lower level processing due to their item design, or that the methodology was insufficient to capture the full range of higher-level processing that occurred in the administration of both tests, due to participants' difficulty in verbalising expansive thoughts that are required for coding of higher level processes.

A third finding is that the number of cognitive processes identified for TOEFL is greater than that for IELTS. As noted in section 4.2.1, there were a greater number of observable strategies identified for TOEFL than IELTS. Participants displayed more instances of observable codable behaviour than when engaging with the IELTS test. A greater number of stimuli in the interviews resulted in a larger number of verbalisations and hence, a greater number of processing codes allocated to TOEFL than IELTS. The number of items is broadly equal in both IELTS and TOEFL, therefore something about the item types inspired more observable behaviour in TOEFL. Many item stems in IELTS were shorter, especially heading, matching information and summary completion items. This allows participants to keep information related to a greater number of options in their working memory while engaging with the text. Options in TOEFL are often very long, as are the question stems. This forced revisits by participants and more strategic actions to manage item completion.

Fourth, the pattern of cognitive processing across the two tests appears to be consistent. The most frequently coded process was 'lexical access' (P2), which reflects the high proportion of visible strategizing by participants. P4c and P4s were the other most frequently-used codes for both tests, whereas P5w, P5s/c and P7 were infrequently used. This suggests that the coding algorithm was applied consistently across the tests. A clear similarity between the tests is the reliance upon lower level processes to access and manage the items and text. Lower level processing is utilised in items designed to access higher level processes as part of test management; participants identify key words in the item stem and options and in the text as a means of identifying relevant parts of the text to engage with. This occurred in TOEFL, despite relevant paragraphs being marked for each item, and in some cases, specific words and sentences are highlighted to minimise the time required to locate relevant sections.

A frequent approach by participants was to identify key words in item stems and options, and then use these to identify relevant areas of the text. This was a consistent pattern for all item types. For the TOEFL test, the second most frequent process was 'establishing propositional meaning at the clause level' (P4c), and for IELTS, 'establishing propositional meaning at the sentence level' (P4s). Coding at the clause level occurred almost twice as frequently in TOEFL as in IELTS, suggesting that the specific unit of analysis required to

respond to the majority of items in the TOEFL test was the clause (subject and verb). For IELTS, sentence level processing (multiple clauses) was proportionally of more significance than in the TOEFL test.

It is also noteworthy that for both tests, the most frequently-used high level process is 'forming a mental model' (P6). This code refers to participants who link propositional content across sentences and paragraphs to form localised understanding about the purpose of what they are reading. There were twenty coded instances of participants performing this in relation to IELTS and thirty-three in TOEFL. This outcome, coupled with the higher percentage of high level processing TOEFL compared to IELTS suggests that TOEFL items are more explicitly designed to elicit high level processes than those in IELTS. To confirm this, the next research question breaks down the findings in relation to individual item types to discern which item types target specific processes and the extent to which the test developers have been successful. The next two sections will reflect on the implications of the findings in this study in relation to wider literature in relation to IELTS (section 4.4.3) and TOEFL (section 4.4.4).

4.4.3. Reflections on the emergent cognitive processes in IELTS

As noted in the literature review, there is a paucity of research related to the cognitive demands associated with both tests. This issue is most pertinent to IELTS, as the reading section has not undergone a significant revision since 1995. Weir et al (2009b) note that both the IELTS handbook (2005) and website do not contain information related to the construct. A new version of the handbook was produced in 2007, although the only changes made to the reading section were cosmetic. The range of item types suggest that IELTS developers adopted a componential approach to reading and in terms of cognitive coverage, must conceive of some overlap between the item types, as there is no uniform approach to the number of each item type in each version of the test. Individual texts may be matched with as few as two item types, and it is possible that a complete test (composed of three texts) will not contain the full coverage of item types used in IELTS. According to the handbook, at least one text will contain "detailed logical argument" (IELTS, 2007: 7), suggesting the content of this text is likely to require higher-level processing to formulate

meaning, although there are no guidelines to suggest that this text is accompanied by specific item types that target higher-level cognitive processing. As a result, the nature of cognitive processing in the IELTS test is blurred. The findings here present evidence of cognitive processing, although the extent to which low and high level processes are associated with specific item types is investigated as part of research question 4. This section will examine the findings in the literature in more depth and compare them to the findings in the present study.

Weir et al (2009b) provide the most comprehensive attempt to account for the cognitive processes embedded in the reading section of the IELTS test, although this study is not successful at identifying higher or lower level processing, or the relative importance of individual cognitive processes in the IELTS test. Their findings can partially be compared with the findings in this study. Tables 2.5 and 2.6 are reproduced below from the literature review for ease of comparability. Table 4.15 highlights specific strategies for specific item types if the strategy use recorded for that item type is above a certain threshold value, giving an overall impression of the importance of each strategy for the test. Strategies are elaborated in table 4.16:

Task Type	Section	Text Prev. + mean >.2			Response Strategy + mean > .15											Locating Information + mean > .3				
		PR1	PR2	PR3	ST1	ST2	ST3	ST4	ST5	ST6	ST7	ST8	ST9	ST10	ST11	ST12	LI1	LI2	LI3	LI4	LI5
MCQ	E1.2	+	+		+	+	+						+	+			+				
	E2.1	+	+		+	+	+						+	+				+			
Sent Comp	E3.1	+	+		+	+					+			+			+				
Summ Comp	E1.3	+	+	+	+	+	+						+	+	+	+		+	+		
	F1.2		+		+	+							+	+				+	+		
Heading	F3.1	+	+			+	+	+					+	+				+	+		
	E2.4		+	+				+							+	+			+		
Locate Info	E2.3	+	+			+	+						+	+				+	+		
	F2.1	+	+			+	+	+					+	+				+	+		
Y/N/NG	E1.1		+		+	+	+						+	+				+	+		
	E2.2		+			+	+				+		+	+		+		+	+		
	F1.1		+			+	+	+					+	+				+	+		
	F2.2	+	+			+							+	+				+	+		
Match	E3.2	+	+			+	+						+	+				+			
	F3.2		+	+	+	+							+	+			+	+			

Table 4.15. Text preview, response strategy and locating information by item type (Weir, et al 2009b: 174)

Sequence of reading activities	Strategies for responding to specific items	Information base for the response
PR1 read the text or part of it slowly and carefully **PR2** read the text or part of it quickly and selectively to get a general idea of what it was about **PR3** did not read the text.	**ST1** match words that appeared in the question with exactly the same words in the text **ST2** quickly match words that appeared in the question with similar or related words in the text **ST3** look for parts of the text that the writer indicates to be important **ST4** read key parts of the text such as the introduction and conclusion **ST5** work out the meaning of a difficult word in the question **ST6** work out the meaning of a difficult word in the text **ST7** use my knowledge of vocabulary **ST8** use my knowledge of grammar **ST9** read the text or part of it slowly and carefully **ST10** read relevant parts of the text again **ST11** use my knowledge of how texts like this are organised **ST12** connect information from the text with knowledge I already have	**L1** within a single sentence **L2** by putting information together across sentences **L3** by understanding how information in the whole text fits together **L4** without reading the text **L5** could not answer the question

Table 4.16. Strategy taxonomy for IELTS reading test (Weir et al, 2009b)

A significant barrier to comparison between the findings in the present study and the findings of Weir et al (2009b) is that there is no readily discernible differences between *processes* and *strategies* in their rubric, which led to the under-representation of the processing core. 'Inferencing', for example, is presented solely as establishing the meaning of difficult words in context, whereas this study has demonstrated that inferential reasoning in a text can occur on multiple levels (word, clause, sentence, across sentences) and their definition does not include deriving the intentions of the author that may not be explicitly stated. There is also ambiguity regarding the point at which putting information together across sentences ('building a mental model') becomes 'text-level representation'. ST1 ('word recognition') and ST2 ('lexical access') clearly relate to specific elements of the processing core, while others (ST9, ST10) could relate to establishing propositional meaning (P4s/c) up to text-level representation (P7). Categories L1-3 appear to relate most specifically to higher and lower level processing, with code L2 specifically outlining that test takers put

information together across sentences to answer these items. This data appears to indicate that the test places greater emphasis on higher level processing than lower level processing, in contrast to the findings in this study, which suggest that the overwhelming basis for responding to items is at the sub-sentence level.

A majority of the cognitive processes identified in this study for both IELTS and TOEFL were low level. With the exception of strategies ST5 and ST6, none of the strategies in Weir et al's rubric explicitly relate to higher-level processing as conceived by Khalifa and Weir (2009). According to Weir et al (2009b), the most commonly reported location for responses was across sentences for all item types, suggesting that most item types are capable of eliciting higher-level cognitive processing. Item types which did not exhibit this trait were *multiple-choice* and *matching* items. However, it is noteworthy from table 4.15 that there is little correspondence between codes L2 (putting information together across sentences) and PR1 (reading the text or part of it slowly and carefully). This suggests that the basis for linking information across sentences by Weir et al's participants is *word identification* in an attempt to locate relevant parts of a text. Therefore, the 'information base' component of Weir et al's rubric cannot easily be used to determine whether higher or lower level processing is used. This demonstrated the weakness of employing a strategy questionnaire in analysing participant response patterns. There is uncertainty if the participants have interpreted the categories in the way that the researchers intended, a weakness which can only be reliably overcome by introducing verbal protocols.

Purpura (1999) argues that a lack of evidence of specific strategy usage does not mean that particular strategies were not used. Researchers must account for the possibility that mental processes may go unreported by research participants. Conversely, research participants may present verbalised accounts that they believe the researcher wishes to hear. Learner awareness of strategy use may also vary across individuals (proficiency). The research design of this study accounts for this shortcoming by providing more detailed or innovative stimuli to maximise the opportunity for research participants to verbalise or otherwise report their thought processes. This study therefore accounts for methodological weaknesses inherent in previous studies in IELTS which have relied upon strategy

questionnaires and presents new insights relating to the construct of reading for academic purposes in the IELTS test that has previously not been available to stakeholders.

4.4.4. Reflections on the emergent cognitive processes in TOEFL

This section specifically considers the findings of the study in light of existing literature regarding the construct definition of the reading section of the TOEFL test, specifically focusing on inferential reasoning. As noted in section 4.2.6, Cohen and Upton (2006) provide the most comprehensive study of task requirements in the TOEFL reading test. The authors collected verbal report data from 32 students who were assigned to complete six reading tasks from the LanguEdge Courseware materials which are designed to prepare test takers for the TOEFL iBT. Successful participants were asked to verbalise how they had completed the tasks. However, like Weir et al (2009b), the authors did not aim to distinguish between strategies and processes. However, one code in their rubric clearly related to inferential reasoning (R26; 'verifies the referent of a pronoun'), providing a basis for comparison with cognitive findings from the present study. Additionally, Jamieson et al (2008) provide significant information regarding the development of the new TOEFL reading section which provides further construct information with which to compare the findings from the current study.

The reading section of the TOEFL iBT was centred on *reading purpose* rather than a hierarchy of proficiency based on cognitive processes such as word recognition, fluency and efficiency (Jamieson et al, 2008 The test developers believed that increasingly complex reading purpose would influence the cognitive demands placed upon test takers, allowing test users to discriminate between stronger and weaker readers (Jamieson et al, 2008: 71). As reading purpose is associated with different reading tasks, these tasks are associated with different task features that act as mediating variables which would influence task difficulty. However, as part of the framework, claims that test takers would be able to understand academic texts were underpinned by three sub-claims as part of the final blueprint (Jamieson et al, 2008: 244). These are 'basic comprehension', 'inferencing' and 'reading to learn'. These can be readily linked to the levels of Khalifa and Weir's cognitive processing core. The details of the sub-claims will be discussed directly in relation to

research question 4, which will elaborate which of the item types include the relevant parts of the processing core which link to the sub-claims of the test developers. This section will explicitly look at the *inferencing* sub-claim to determine whether the nature of inferencing identified in this study coheres with that of the developers (Jamieson et al, 2008) and independent researchers (Cohen and Upton, 2006).

The inferencing sub-claim states that successful test takers are able to comprehend ideas that are not explicitly stated in a text, and that this skill that can be distinguished from basic comprehension (Jamieson et al, 2008: 242). Cohen and Upton's (2006) definition of inferencing includes three components in their initial rubric, reproduced here:

R26 Verifies the referent of a pronoun.

R27 Infers the meanings of new words by using work attack skills: Internal (root words, prefixes, etc.).

R28 Infers the meanings of new words by using work attack skills: External context (neighbouring words/sentences/overall passage).

(Cohen and Upton, 2006: 35).

This demonstrates that the working definition of inferencing in Cohen and Upton (2006) has been superseded by the newer definition in the publicly-available research from ETS. The distinction between word-level inferential reasoning from Cohen and Upton (2006) and sentence-level inferential reasoning from Jamieson et al (2008) was maintained in the current study. However, two of these definitions, R27 and R28, recorded only low usage across all item types in Cohen and Upton (2006). Only R26 recorded 'high' use, but in relation to a single item type – 'basic comprehension pronoun reference' (BC-pr; table 4.8), which was not intended by the developers to test high-level inferential reasoning. Data from Cohen and Upton (2006) therefore suggests that inferential reasoning is under-represented in the TOEFL test. This is explicitly the case if we consider the definition of inferential reasoning offered by the test developers mentioned in section 4.3.2:

> "can comprehend an argument or idea that is strongly implied but not explicitly stated in the text, identify the nature of the link between

specific features of exposition and the author's rhetorical purpose, and understand the lexical, grammatical and logical links between successive sentences in a passage" (Jamieson et al, 2008: 244).

This definition offered by the developers focuses on ideas or arguments; propositional content within a text, and following internal logic across sentences. This definition therefore centres on deriving implicit conclusions from multiple propositions. Cohen and Upton centre their definition of inferencing on identifying the meaning of unknown lexis. Therefore, the definition of inferencing in their study only encompasses word-level inferencing, not clause or sentence-level inferencing cited by the test developers. In contrast, data from the current study suggests that both conceptions of inferential reasoning have a small, but significant presence in the TOEFL test (4.49 percent of total codes). Seventeen instances of inferential reasoning were coded in total; seven at the word level (P5w) and ten at the clause or sentence level (P5s/c). This demonstrates that it is possible to discern between inferential reasoning at the word/clause/sentence level in the TOEFL test and provides direct evidence of the inferencing sub-claim by the test developers. This is also further evidence of the success of the methodology in capturing these important instances of inferential reasoning. The next research question will specifically outline which item types revealed evidence of inferential reasoning.

4.4.5. Summary of findings relating to research question 3

In summary, five major claims emerged in relation to the data which directly address the research question *'do these processes reveal differences in IELTS and TOEFL iBT test developers' understanding of the construct of interest?*' First, as understood from the perspective of Khalifa and Weir's (2009) cognitive processing model of reading, each of the tests contain items which measure eight of the nine levels of the model which have been identified for the purposes of this study. This suggests that the two tests have a broadly similar conception of the construct of reading. Second, both tests appear to overwhelmingly target low level processing as defined by the model, with coverage of high level cognitive processes accounting for less than twenty per cent of the overall codings used with participant verbalisations. Third, another similarity between IELTS and TOEFL is that in terms

of high level processing, 'establishing a mental model' (P6) was the most frequently-used. This suggests that some of the higher-level processes in Khalifa and Weir's model are under-represented in both IELTS and TOEFL, including inferential reasoning and forming a text-level representation. Both types of inferential reasoning identified in this study (word and sentence/clause level) are identified in both IELTS and TOEFL, although there were more explicit instances of these cognitive processes in TOEFL than IELTS. There was no previous evidence in the literature of sentence/clause level inferential reasoning for either test, which this study provides. Fourth, the use of Khalifa and Weir's model also revealed that both tests have a truncated conception of the construct. Neither test requires test takers to form an understanding of more than one text and then use this information to answer any items. Code P8 (intertextual representation) was not used for either test. Fifth, TOEFL recorded a greater number of cognitive processing codes than IELTS, in line with the greater number of observable moments of engagement with the TOEFL test than the IELTS test by the participants. The findings in this section represent a significant addition to our understanding the constructs of reading for academic purposes embedded in IELTS and TOEFL.

The next section addresses research question 4. This section will explicitly identify the composition of cognitive processes in terms of different item types. This will address the question of whether specific cognitive processes are associated with particular item types, and in the case of TOEFL, whether the item types elicit the types of cognitive processes that the test developers claim they do. This section includes example verbal evidence to demonstrate which processes are associated with which items.

4.5. Research question 4: Are cognitive processes associated with specific item types? Do individual item types target specific processes or do they elicit a range of processes?

Section 4.5 directly addresses research question 4. This section begins by providing evidence of cognitive processing associated with individual task types in the form of analysis of participant verbalisations. Sections 4.5.1 and 4.5.2 present evidence for cognitive processes

for each of the item types in IELTS and TOEFL respectively. These sections present direct evidence for the claims of cognitive processing in relation to specific task types. Each item type from both IELTS and TOEFL is addressed in an independent section. Each section includes the most frequently-used strategies by the participants for that item type. These are the observable interactions of participants with the items identified in relation to research question 1. Each item type contains a table with the most frequently-used strategies for that item type, and a graph and table depicting the most frequently-used cognitive processes from Khalifa and Weir's (2009) model which are inferred from verbal data produced by participants. Each section also includes example item types from the tests used and a subsequent discussion of participants' engagement with the tasks and the evidence base from which relevant cognitive processes are inferred.

As in previous sections, the strategies and cognitive processes identified are presented as ratios – the number of strategies and cognitive processes identified for each item type is divided by the number of items of that type. This is interpreted using the bands identified in the methodology (table 3.14, section 3.3.4.3). Only strategies which recorded a rating of 'medium' or higher in relation to each item type are included in each section. This gives an indication of *relative* importance of each cognitive process to each item. The verbal evidence relates explicitly to individual item types in the two tests used in this study. For this reason, it is recommended that readers familiarise themselves with the two tests used in the data collection before reading this section. The second part of this section scrutinises the degree of congruence between the outcomes in the research and the claims regarding individual item types made by IELTS and TOEFL test developers and provides reflection on the nature of the cognitive processes embedded in specific item types.

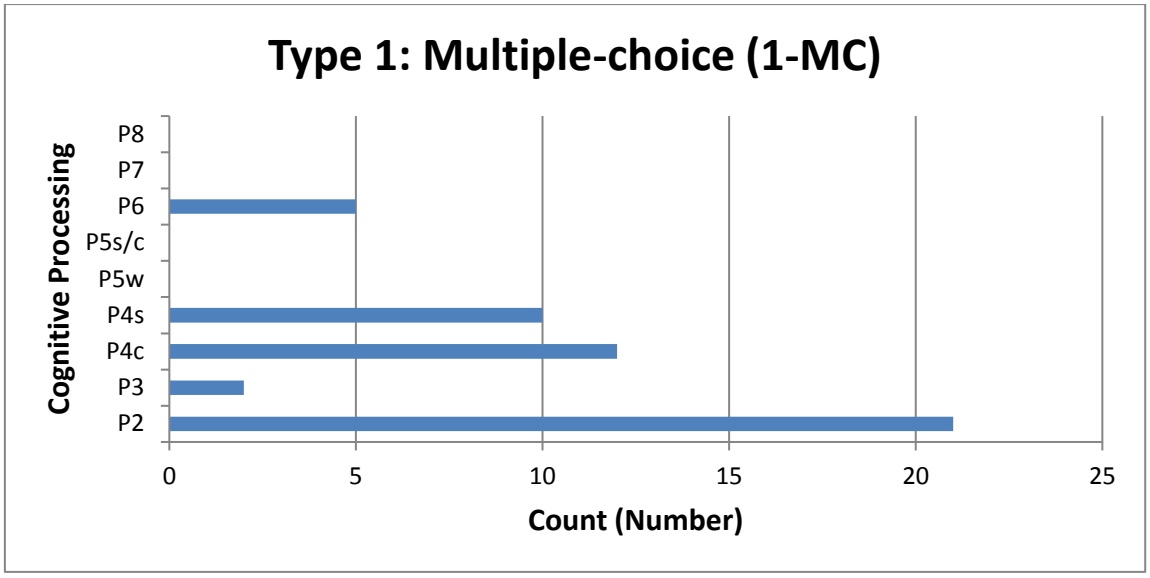
Section 4.5.3 collates the findings from the individual sections and directly addresses research question 4. This section uses a cognitive processing matrix inspired by Buck (2001) to present the findings, to compare the distribution of emergent cognitive processes in the two tests according to item type. This section also includes a critical reflection on the findings in relation to existing literature regarding cognitive processing in the two tests and specifically highlights the contribution to knowledge that the investigation of cognitive processes has added to the literature.

4.5.1. Cognitive processing in IELTS item types

4.5.1.1. Type 1: Multiple-choice (1-MC)

Strategy code	Strategy label	Freq. rate
6	Marks/notes key noun phrase(s) in the questions stem or options	8.00
12	Marks/notes key noun phrase in the text during careful reading	5.00
11	Careful local reading (text)	4.67
13	Marks/notes key verb phrase in the text during careful reading	2.33
18	Returns to the question for clarification: rereads question and/or options	2.33
25	Eliminates option(s) (contradicts key noun/verb phrase/nonsensical response)	2.00
5	Reads question stem(s) and/or option(s) carefully	1.67
2	Identifies the purpose of the question	1.00
9	Marks/notes key prepositional phrase(s) in the question stem or options	1.00
15	Marks/notes key prepositional phrase in the text during careful reading	1.00
19	Searches for key word/phrase (text)	1.00
22	Identifies content-based parallel between paragraphs	1.00
27	Compares question stem/option to a portion of the text	1.00
4	Identifies grammatical or content-based parallel between items/options	0.67
14	Marks/notes key adjective phrase in the text during careful reading	0.67
20	Skimming part of the text for general understanding (expeditious reading)	0.67
21	Marks/notes key phrase in the text during expeditious reading	0.67
24	Identifies paraphrase within text or between text and item stem	0.67
26	Eliminates option (s) (no information found)	0.67
1	Reads question(s) before proceeding to the text	0.33
7	Marks/notes key verb phrase(s) in the question stem or options	0.33
10	Marks/notes key adverbial phrase(s) in the question stem or options	0.33
23	Identifies lexical parallel between parts of the text (matches words across paragraphs or heading)	0.33
28	Checks/confirms/considers option choice after reading portion of text	0.33

Table 4.17. Most commonly-used strategies for IELTS multiple-choice items



	P2	P3	P4c	P4s	P5w	P5s/c	P6	P7	P8
Total processes	**21**	**2**	**12**	**10**	**0**	**0**	**5**	**0**	**0**
Processing ratio	**7.00**	**0.67**	**4.00**	**3.33**	**0.00**	**0.00**	**1.67**	**0.00**	**0.00**

Table 4.18. Most frequently-identified cognitive processes for IELTS multiple-choice items

Sample item

Questions 1—3
*Choose the correct letter, **A**, **B**, **C** or **D**.*
Write the correct letter in boxes 1—3 on your answer sheet.

1 The main topic discussed in the text is

A. the damage caused to US golf courses and golf players by lightning strikes.
B. the effect of lightning on power supplies in the US and in Japan.
C. a variety of methods used in trying to control lightning strikes.
D. *a laser technique used in trying to control lightning strikes.

Textual reference:

Striking Back at Lightning with Lasers

Seldom is the weather more dramatic than when thunderstorms strike. Their electrical fury inflicts death or serious injury on around 500 people each year in the United States alone. As the clouds roll in, a leisurely round of golf can become a terrifying dice with death — out in the open. A lone golfer may be a lightning bolt's most inviting target. And there is damage to property too. Lightning damage costs American power companies more than $100 million a year.

The IELTS test used in this research contains three multiple-choice items (questions 1-3). These items were completed by participants 1 and 4. Participant 1 answered all three items correctly (selecting options D, A and A respectively), whereas participant 4 answered item 1 incorrectly and items 2 and 3 correctly (selecting options C, A and A respectively). Multiple-choice items in IELTS are composed of a stem and four options. The stem and key form one grammatically complete, single clause sentence. All options are designed to be grammatically possible and relate to the content of the text. A wide range of strategies were employed in the completion of these three items. Six strategies were rated 'very high' based on observations of participants. These frequently used strategies suggest a common approach to these items of identifying key words, identifying their equivalents in the text and then reading around them to attempt to establish propositional meaning. Options were eliminated if they did not represent an accurate meaning of the text.

Evidence of a range of cognitive processes emerged in relation to multiple choice item types, positive evidence that this item type can target both higher and lower level processes. However, higher level processes were not restricted to those items which state they require participants to form a mental model of the text in their mind, such as item 1. This item asks participants about the main topic of the text. Both participants addressed this question using lower level processes by engaging with topic sentences and key words in the title. Higher level processes emerged in relation to item 3, which asks participants about detailed information in the passage, supposedly prioritising careful local reading and establishing propositional meaning. However, the key for item 3 is option A, which as a sentence, requires participants to link information across sentences and to infer the meaning of 'support' in financial terms ('receive funds'), which is unstated in the text. Both participants 1 and 4 exhibited this higher order processing in relation to item 3, suggesting it was crucial to item completion.

For item 3, participant 1 quickly identified 'University of Florida' as a key phrase. She also underlined an extended verb phrase in the text, linking this to the noun phrases. After checking the aim of question 3 once more, she then identifies the second university in paragraph 5 (University of New Mexico). Question 3 requires the test taker to draw parallels between these two paragraphs. She highlights 'University of New Mexico' in paragraph 5.

She then returns to paragraph 3 and identifies 'backed' and 'support', recognising their synonymy, and recognises the same company mentioned in paragraphs 3 and 5 (multiple examples of lexical access). She then skims paragraphs 3 and 5, while considering options A and B ('source' versus 'techniques'). She chose option A on the basis of 'support' and 'backed' being synonymous with 'receive funds' in option A. Despite this item being overtly related to detailed information in contrast to item 1, there is evidence that the participant engaged in a high level of cognitive processing. Knowledge of synonymy is important for this item as the text does not explicitly state that both universities receive funding from the EPRI, (only 'backed' and 'supported').

> *"'Source'? It's the same as 'support' and 'backed up'. I think at this time, 'back up' and 'support', I think maybe they are both from the same company, or other departments, then I want to find another information about the techniques or other things, so I go back to here* [from paragraph 5 to 3] *to reread it. I think these two information are enough for me to choose A. I think they are the same institute, and they [Florida and New Mexico] all gain support [from 'EPRI'] but maybe I have done the wrong answer, because I didn't find the 'funds', just 'support' and 'back up', but maybe from different ways, instead of funds, maybe from other things"* [Participant 1, test 1, item 3].

This is evidence of building a mental model, as participant is explicitly required to recognise that 'support' and 'backing' can take the form of funding and had to draw on multiple-clause propositional content from two paragraphs to select this option. Participant 4 also displayed an ability to infer information ('backed up' and 'support'). This was not coded as 'inferencing at the text level', as the research mandate stipulated coding at the highest level of the model possible. As participants demonstrated linking propositional content across paragraphs, this item was associated with the need to form a mental model.

Participant 1 also displayed higher level processing in completing item 1; a clear intention of the item design which targets the main topic of the text. Participant 1 completed item 1 after she had completed question 3 and therefore had gained some knowledge of text

organisation. She returned to question 1 to refresh her working memory of this item. She then reread portions of paragraphs 1 and 2 carefully. She returns to question 1 to consider the options and eliminated options A and B. She explains that these options refer to details in the text (golf players and the effect of lightning of the US and Japan). This is evidence of higher-level processing and that she is forming a mental model of the text. Considering options C and D, she notices the difference is that one represents a single technique and the other option a variety of techniques. She eliminated option C by scanning to see how many techniques were mentioned, but found nothing, so selected option D, further evidence of building a mental model:

> *"I think this one and this one [Options A and B] is not the main topic, it's may be just some detailed information included in the passage. I think this kind of information could not be the main topic of the passage, so I delete these two. The other two, the difference between these two is 'a variety of methods' [Option C] and 'a technique' [Option D], so I want to find if there [is] included any information to indicate there are a lot of methods, through scanning maybe, but I can't find it, so I chose this one [Option D]"* [Participant 1, test 1, item 3].

Evidence of the importance of higher level processing to this item type can be found in the response of participant 4. She answered this item incorrectly. She forms a mental model of the first two paragraphs. She underlines a phrasal verb ('hit back') from the first sentence of paragraph 2. This only coheres as a key phrase in the sentence in relation to the content of paragraph 1. She demonstrates understanding of this term in context and immediately understands that the text will be about controlling lightning, with this paragraph offering a solution. It is clear that she has built a mental model by integrating new information from paragraph 2 with the mental representation she has established of paragraph 1:

> *"The first paragraph talks about damage of the lightning, and paragraph 2 began to talk about the method to control the lightning or use the lightning. There, I can figure out the main point of this passage, it's about*

several ways to control the lightning, not one, so I can choose the option 3 [C]" [Participant 4, test 2, item 1].

She most likely answered incorrectly by failing to form a mental model of the text beyond the first two paragraphs, and demonstrating insufficient knowledge of the text structure by claiming that the purpose of the text would be identified in paragraph 2. The correct answer is Option D ('a laser technique used in trying to control lightning') rather than Option C ('a variety of methods'). Nonetheless, she demonstrates propositional understanding of the parts of the text that she had parsed carefully, but this is not sufficient to answer the item correctly.

In contrast to higher level processes associated with items 1 and 3, participant 1 responded to item 2 with a greater emphasis on lower level processing. Her initial strategy was to go through the items highlighting key words before moving to the text:

"I want to find some information about damage, so I underline it… and the 'courses' and 'golf players'… The word 'effect' is the main information and 'in US and Japan'… Option C, methods… yeah, this one I want to find some methods… to do what?" [Participant 1, test 1, item 1].

The participant read the three multiple-choice questions before proceeding to the passage. She immediately underlined 'main topic' of item 1 then turned her attention to the title of the passage and underlines 'lightning', to determine whether the title offers any clues to the main topic discussed in the article (the purpose of Q1) (lexical access). She then went through the four options for question 1, highlighting key words, all noun and prepositional phrases. The purpose of this strategy was to memorise the key terms which may appear in the text or be paraphrased. She was able to explain the links between the words that she is highlighting. For options C and D, the participant also noticed that the same wording and sentence structure is used. She therefore hesitated when reading option D to establish the difference in meaning between the two options Option D highlights one technique for controlling lightning, whereas option C states a 'number of techniques' (establish propositional meaning at the clause level).

The participant frequently returns to the questions for clarification, as there is much information to hold in her working memory (stem and four possible options for each item). She demonstrates word matching (lexical access) as she underlines words in paragraph 1 that match those used in the options for questions 1 and 2 ('damage' and '500 people') and reads around them. She also circles the word 'alone', demonstrating that she is able to hold information from at least three options from question 2 in her working memory. She realises that the phrase 'the United States alone' contradicts option C in question 2, so returns to the question to eliminate it. The answer to question 2 was reached by word matching ('damage') and by paraphrasing ('buildings' and 'property') (primarily lexical access):

> *"'Property', I think maybe this is the paraphrase of 'buildings' [Option A], so I think it might be important. So [at this moment] I chose 'A'"*
> [Participant 1, test 1, item 2].

Participant 4 answers item 2 correctly, also by focusing on propositional meaning of key and distractors. This was achieved largely by a process of elimination. Key phrases from the alternative options were contradicted in the text. She effectively summarises the content of paragraph 1, then underlines the number of people killed each year by lightning. She eliminates option C on the basis of this information as it contradicts the adverbial clause identified in the text. She also eliminates option D as it contradicts information contained within the final sentence of paragraph 1. These two actions and associated verbalisations highlight the reasoning behind separating code 4 into two (clause and sentence level propositions). The key contradiction between the text and option C occurs at the clause level, whereas in option D, this occurs at the sentence level. The proposition is different, whereas in option C, the difference in meaning is contained within the adverbial clause. Option B is eliminated as the participant correctly surmises that this option refers to a detail in the text. She checks this response by returning briefly to the text to confirm her choice:

> *"At this moment, I have read this paragraph, but I didn't find any useful information maybe I can use to answer the questions. I have*

deleted these two [Options B and C] but there is still two I haven't decided, so I go back to find out which one is maybe not sure, so I go back… maybe when I read this one, I want to find some information but I didn't find it, so I go back to paragraph 1; I want to get some useful information that I can use, so I reread it [paragraph 1]. 'Property', I think maybe this is the paraphrase of 'buildings' [Option A], so I think it might be important. So [at this moment] I chose 'A'" [Participant 4, test 2, item 2].

Participant verbal records of these items are crucial pieces of evidence regarding the higher and lower level processes elicited by the items. The data suggests that multiple choice items in IELTS can elicit both higher and lower level processes. A further finding is that to elicit a mental model of the text from test takers, test developers and item writers must compose the distractors of detailed points from the text, in contrast to the key, which will summarise the argument of the text in such a way that it cannot be selected on the basis of content in the title or opening sentence. Thus, cognitive evidence of why participants eliminated options is as important as cognitive evidence of why participants selected the key for this item. The title is a close approximation of option D, but if the aim of the developers is for test takers to only cohere the key with the title, then this item cannot be said to target higher level processes. This only occurs when test takers demonstrate engagement with distractors also. Neither participants 1 or 4 stated that the title was a crucial piece of evidence in responding to this item, indicating that this would have been insufficient evidence for them to answer confidently.

4.5.1.2. Type 2: Identifying information (Yes/No/Not given) (2-IDi)

Strategy code	Strategy label	Freq. rate
11	Careful local reading (text)	**2.67**
18	Returns to the question for clarification: rereads question and/or options	**2.67**
6	Marks/notes key noun phrase(s) in the questions stem or options	**1.67**
12	Marks/notes key noun phrase in the text during careful reading	**1.33**
13	Marks/notes key verb phrase in the text during careful reading	**1.00**
27	Compares question stem/option to a portion of the text	**0.67**

9	Marks/notes key prepositional phrase(s) in the question stem or options	**0.33**
19	Searches for key word/phrase (text)	**0.33**
20	Skimming part of the text for general understanding (expeditious reading)	**0.33**
1	Reads question(s) before proceeding to the text	**0.33**
8	Marks/notes key adjective phrase(s) in the question stem or options	**0.33**
30	Hesitates while answering to reconsider choice	**0.33**

Table 4.19. Most commonly-used strategies for IELTS Identifying information (Yes/No/Not given) items

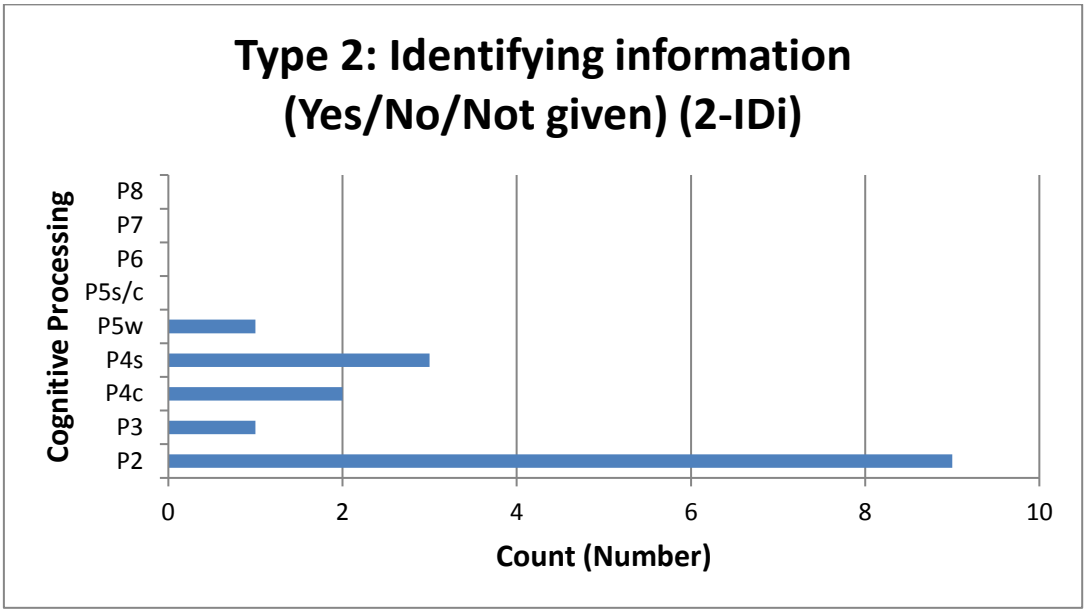


Figure 4.4. Frequency of cognitive processes for IELTS Identifying information (Yes/No/Not given) items

	P2	P3	P4c	P4s	P5w	P5s/c	P6	P7	P8
Total processes	**9**	**1**	**2**	**3**	**1**	**0**	**0**	**0**	**0**
Processing ratio	**3.00**	**0.33**	**0.67**	**1.00**	**0.33**	**0.00**	**0.00**	**0.00**	**0.00**

Table 4.20. Most frequently-identified cognitive processes for IELTS Identifying information (Yes/No/Not given) items

Sample item

Questions 11—13

Do the following statements agree with the information given in Reading Passage 1?
In boxes 11—13 on your answer sheet write

YES *if the statement agrees with the claims of the writer*
NO *if the statement contradicts the claims of the writer*
NOT GIVEN *if it is impossible to say what the writer thinks about this*

11. Power companies have given Diels enough money to develop his laser. _____ **[NO]**

Textual reference: "Bernstein says that Diels's system is attracting lots of interest from the power companies. But they have not yet come up with the $5 million that EPRI says will be needed to develop a commercial system…"

There are three examples of 'identifying information' (yes/no/not given) in the IELTS test used in this study (items 11-13). This item type presents participants with a series of statements relating to different parts of the text. Participants are required to identify whether these statements are given in the text (yes), are contradicted by the text (no) or are unrelated to text content (not given). Items 11-13 relate to text 1, so formed part of the first test, so were completed by participants 1 and 4. This item type proved more difficult for the participants than the multiple choice items. Participant 1 did not record any correct responses. Participant 4 recorded 2/3 correct responses (items 11 and 13, answering item 12 incorrectly). This is problematic from the perspective of attempting to identify the crucial cognitive processes for completing this item type. If participant 1 did not answer any items correctly due to not processing text at the appropriate level, then the emergent cognitive profile for this item type could be under-represented at the higher levels of the model. However, participant 4 answered two items correctly, therefore the contrast between responses to items 11 and 13 by participants 1 and 4 will be used to identify the necessity of higher-level processing with this item type. Evidence of strategy use was sporadic in relation to this item type. Two strategies recorded 'very high' use. Three additional strategies recorded 'high' use. These strategies in combination suggest that participants were focused on establishing propositional meaning either at the clause or sentence level. The frequency that participants returned to the question stems indicate that they were required to engage carefully with multiple linguistic elements in the question stems.

The high proportions of key words (nouns and verbs) that were highlighted explains why the only cognitive process to record 'very high' use was 'lexical access' (P2). Participants used key words to identify parts of the text that were relevant to the item stems before reading those parts in depth. Few participants, coupled with a low facility rate (2/6 for both

participants combined) resulted in limited data for this item type. Only one instance of higher-level processing was recorded – there is limited evidence of one participant inferring information at the word level. Participant 4 displays evidence of identifying key words as her principle strategy of identifying relevant parts of the text:

> *"I'm scan[ning] the questions… Question 13 mentioned some new[ly] nouns, 'weather forecasters'. The paragraph did not mention [these] before. So, I need to pay attention to these words after [when] they occur"* [Participant 4, test 2, item 13]

She answers item 13 correctly, relying on key nouns to identify the correct part of the text. The correct response was 'not given' and she based her selection of this option on the use of lexical access to locate the appropriate part of the text. The principle clue was the noun phrase 'weather forecasters', which she identified in the stem and locates a parallel phrase in paragraph 3:

> *"I'm trying to find some[thing] related to the weather forecasters and it's about this paragraph is about the implications of the new methods, and this 'laser thunder factory', I found a function to use [the laser] but I didn't find weather forecasters, and I mean the paragraph 3 does mention about forecasting the weather, but not to… but not the implication and it didn't mention forecasters, just 'forecasting the weather'"* [Participant 4, test 2, item 13].

The participant demonstrates that she identified a portion of the text that highlights a use of 'the laser', also when she underlines the verb 'shake' in the final paragraph at 9.30, she demonstrates that she has established propositional meaning at the clause level at this point. She is also able to differentiate between 'forecasting the weather' and identifying in paragraph 9 that no weather forecasters themselves are cited as interested in the development. This clearly indicates that participants should not be able to select 'not given' in the text by identifying key noun phrases in the stem of the item and confirming that they are not present in the text. For 'not given' items, key phrases should be repeated or

paraphrased in the text, encouraging participants to identify propositional meaning of relevant utterances to determine whether the statement in the item stem is consistent with information presented in the text.

Participant 4 answers item 11 correctly. The correct response to this item is 'no'. The participant correctly establishes propositional meaning at the sentence level by drawing a parallel between the item stem and information across sentences 1 and 2 in paragraph 8, with some inferencing at the word level as she links a pronoun in sentence 2 to its referent in sentence 1:

> "It's about 'enough money'. And the question 11, the main point is if he has enough money, and 'he has not yet come up with' means that they have not enough money" [Participant 4, test 2, item 11].

The participant again identifies key terms ('enough money') and relates them to a relevant part of the text. In this instance however, she cites evidence from two sentences to arrive at the correct response. The relevant sentence in paragraph 8 identifies the relevant agents stated in the previous paragraph by pronoun only and the scientist is not directly referred to in the sentence. This is the only example of higher level processing that was evident from participant verbalisations in relation to this item type. In contrast, participant 1 answers this item incorrectly. She initially identified the relevant portion of the text ('they haven't yet come up with $5 million') and correctly summarises that this quote contradicts the question stem. She goes on to recognise that it is possible for the researcher to obtain funding ('maybe they can receive the money'), but by focusing on the key words from the question stem, there is no evidence that the participant parsed the present perfect tense structure ('have given'). Incorrectly parsing the item stem may be the basis of the participant's failure in relation to this item.

Evidence from participant 4 completing these items suggested that she did not complete these items independently. She reads all three question stems before attempting to identify the relevant portions of the text. This is seen when she uses lexical access to identify and underline 'lots of interest', at 7 minutes, 33 seconds (paragraph 8, line 1), indicating that she

is holding information from item 13 in her working memory before she has answered items 11 or 12. She also returns to the items for clarification several times, identifying key terms in each question before returning to the text, to provide key terms to scan for. Test-wiseness in the form of knowledge of the test format also guides her approach to the text, as she is aware that items progress in the order that information appears in the text ("according to IELTS, the questions may be after one paragraph, one paragraph [i.e. questions progress according to the text]", participant 4, test 2, items 11-13). She uses this knowledge to identify relevant parts of the text to scan for key terms that she has identified.

Both participants 1 and 4 respond incorrectly to item 12. Participant 4 selected 'not given' on the basis that the words 'real storms' are not actually mentioned in the text, so she believes that the statement is not relevant to the text *("Because the key words in 12* [are] *'real storms', but in the text they are not mentioned"*, participant 4, test 2, item 12). Hypothetically, selecting the correct response ('Yes') requires the reader to parse sentence 4 in paragraph 8 ('…forthcoming field tests will be the turning point…'); to recognise the synonymy between 'real storms' and 'field tests' and 'turning point' to represent a change in fortune in relation to the previous lack of funding. This suggests that the developers intended for participants to possess knowledge of complex lexis and to integrate this information across three sentences. The participant selected 'No on the basis of cognitive processing at the lexical level, indicating that this is insufficient level of processing to respond correctly. She did not present evidence of processing at a higher level for this item. Participant 1 does not answer any of the items in this section correctly, despite identifying the correct portions of the text for each of the items, reasoning through the text and demonstrating knowledge of the discourse structure and propositional meaning of the text. The participant's failure in these items is therefore due to an inability to process at a sufficiently high level, indicating that the developer targeted high level processes in this item type, although only one instance can be verified from participants' responses from the current research.

4.5.1.3. Type 2: Identifying information (True/False/Not given) (2-ID)

Strategy code	Strategy label	Freq. rate
6	Marks/notes key noun phrase(s) in the questions stem or options	1.75
7	Marks/notes key verb phrase(s) in the question stem or options	1.38
33	Writes note/labels part of text/item	1.13
11	Careful local reading (text)	0.50
18	Returns to the question for clarification: rereads question and/or options	0.50
13	Marks/notes key verb phrase in the text during careful reading	0.50
27	Compares question stem/option to a portion of the text	0.50
8	Marks/notes key adjective phrase(s) in the question stem or options	0.50
28	Checks/confirms/considers option choice after reading portion of text	0.50
12	Marks/notes key noun phrase in the text during careful reading	0.38
24	Identifies paraphrase within text or between text and item stem	0.38

Table 4.21. Most commonly-used strategies for IELTS Identifying information (True/False/Not given) items

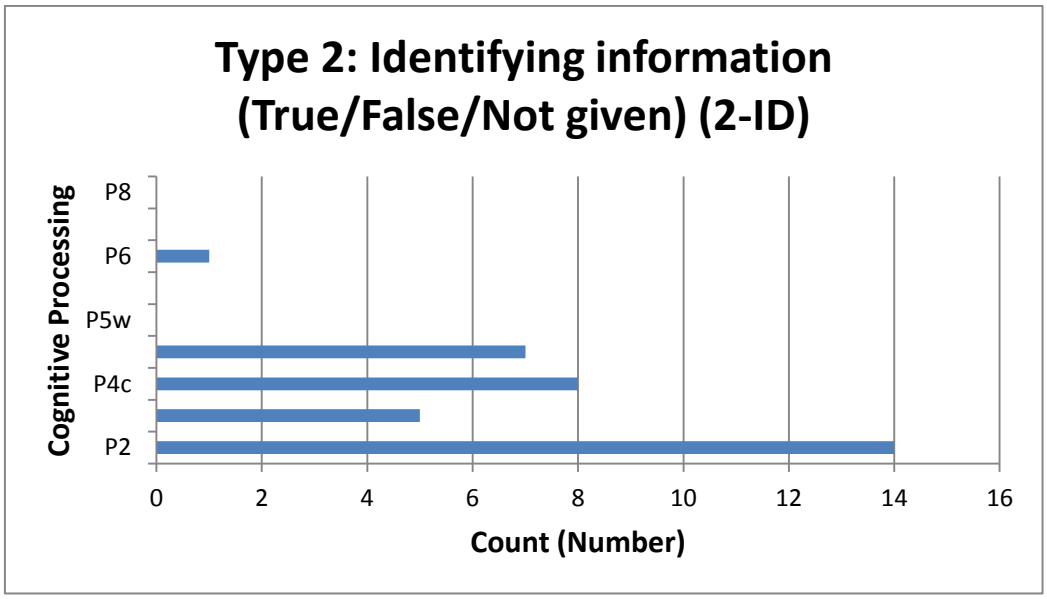


Figure 4.5. Frequency of cognitive processes for IELTS Identifying information (True/False/Not given) items

	P2	P3	P4c	P4s	P5w	P5s/c	P6	P7	P8
Total Processes	14	5	8	7	0	0	1	0	0
Processing ratio	1.75	0.63	1.00	0.88	0.00	0.00	0.13	0.00	0.00

Table 4.22. Most frequently-identified cognitive processes for IELTS Identifying information (Yes/No/Not given) items

Sample item

Questions 19—26

Do the following statements agree with the information given in Reading Passage 2?
In boxes 19—26 on your answer sheet, write

TRUE	*if the statement agrees with the information*
FALSE	*if the statement contradicts the information*
NOT GIVEN	*if there is no information on this*

19 Nineteenth-century studies of the nature of genius failed to take into account the uniqueness of the person's upbringing. _____ **[FALSE]**

Textual reference: "The nineteenth century saw considerable interest in the nature of genius, and produced not a few studies of famous prodigies. Perhaps for us today, two of the most significant aspects of most of these studies of genius are the frequency with which early encouragement and teaching by parents and tutors had beneficial effects on the intellectual, artistic or musical development of the children…"

The IELTS test used in this study contains eight examples of this item type (text 2, items 19-26). These items were completed by participants 2 and 5. Participants are presented with a series of statements and are required to determine whether they agree with information in the text (true), contradict information in the text (false) or are statements reflecting information not contained within the text (not given). The possible responses to this item type indicates that statements are factual, and will be best used in relation to factual texts. The statements vary significantly in their length and structure. They may be simple, single clause sentences, or more complex multi-clause sentences, modified by prepositional phrases or participle clauses. Test takers may be directed to specific arguments given by people named within the text. Statements may refer to information contained within sentences, within paragraphs or across paragraphs. Participants may determine that there is no textual reference for 'not given' statements at all.

Participant 2 recorded four items correct out of the eight (items 20, 21, 25 and 26 were answered correctly). Participant 5 responded to all items correctly with the exception of item 19; therefore, no record of a correct response was available for item 19. Participants employed a variety of strategic actions to complete this item type, with no item type

recording 'very high use'. Three strategies recorded 'high' use which indicate that the participants used key concepts to identify relevant parts of the text. The item stems appeared to place a heavy cognitive load on participants. These strategies indicate that participants did not complete these items in isolation; they tended to identify preliminary responses to the item stems based on their memory of the text they gained from engaging with the text to answer previous items. This is evident in the behaviour in participant 2, who responds to all eight items within the first 2 minutes and 15 seconds of commencing this section. She responds correctly to four of those eight items within that time. This is evidence that she is using her memory of the parts of the text that she has engaged with up to that point, as the probability of guessing four out of eight answers correct in three-option multiple-choice items (true, false, not given) is just under 17%[5] (determined using the probability mass function of the binomial coefficient [6]). This is further backed up by participant 2's verbalisation in relation to item 19:

> *"I wrote 'true' based on my memory. I saw the word from the text, the passage; 'upbringing'. The meaning of this item is similar to, correlated to what I read [at] the bottom of paragraph 3; I want to make some judgments according to my fresh memories. "Failed to take into account of the uniqueness of the person's upbringing" … that is what I interpreted; the parts of paragraph 3, the bottom part; 'upbringing'"* [Participant 2, test 1, item 19].

Further evidence of answering this item type holistically is given by participant 5. After reading the first 9 lines of paragraph 3, participant 5 returns to the items and re-reads items 19, 20 and 21. She underlines key terms in all three items before returning to paragraph 3. The participant verbalises that items 19 and 20 refer to failings with the studies conducted in the 19th century *("Question number 19 and 20, they said the problems of this kind of research* [Participant 5, test 2, items 19-20]"); also the words 'failed' and lacked' in the stems of items 19 and 20 are highlighted almost simultaneously. She then returned to the

[5] 0.16939

[6] $\Pr(X = k) = \binom{n}{k} p^k (1-p)^{n-k}, where \binom{n}{k} = \frac{n!}{k!(n-k)!}$, where *n* = number of items, *k* = number of items correct, *p* = probability of correct response by guessing (1/3).

text and circled the word ‘difficulty’, recognising that this portion of the text may offer evidence in relation to items 19 and 20 (lexical access).

No record of a correct response to item 19 was obtained from participants 2 and 5. Participant 2 cites the word ‘upbringing’ cited in line 10 in paragraph 3 and ‘the bottom part’ of the paragraph. However, this part of the text only indicates that nineteenth century studies were unscientific, not that they were unconcerned with upbringing. The first half of the paragraph clearly indicates that nineteenth century studies took upbringing into account (*“two of the most significant aspects of most of these [nineteenth century] studies of genius are the frequency with which early encouragement and teaching by parents and tutors had beneficial effects on the intellectual, artistic or musical development of the children”*). This suggests that participant 2 focused on the repetition of one key word in the stem (‘upbringing’) to the detriment of others (‘nineteenth century studies’) which would have directed her to the beginning of the paragraph. A parallel that she fails to draw is the phrase ‘uniqueness of a person’s upbringing’ (item stem) and link this directly to line 11 of paragraph 3: “how common or exceptional they were”. This is evidence that the participant was able to establish propositional meaning at the clause level, but not at a higher level of processing. As a result, she selects the incorrect response. The verbalisation

> *“they just describe anecdotes, single anecdotes without taking into account the context”* [Participant 2, test 1, item 19].

demonstrates that she was able to parse and establish meaning, but not relate it to the noun phrase in the sentence stem in item 19.

Participant 5 also responded incorrectly to item 19, also answering ‘true’ suggesting a similar focus on the wrong part of the paragraph. However, participant 5 responded quite differently to this item. She correctly identified the key term ‘nineteenth century’ as important, then moved to paragraph 3. She reads the opening sentence and identifies ‘two significant aspects’ and labels them ‘1’ and ‘2’ respectively. This is labelled ‘syntactic parsing’ rather than ‘establishing propositional knowledge as she does not at this point demonstrate that she has parsed the information contained within the sentence, but has

used the sentence structure to identify where the information regarding the two aspects are located. Describing her thought process approximately 90 seconds into the video, she states that she *'did not fully understand'* (Participant 5, test 2, item 19) the propositional meaning, which suggests that her processing of this part of the text was insufficient to respond correctly, so therefore used her knowledge of sentence structure to highlight the key noun/verb phrases in lines 3-6.

Therefore, it can be conclusively stated that item 19 targets cognitive processing at a higher level than 'syntactic parsing' and 'establishing propositional meaning at the clause level'. The question is designed to activate test takers' memory of several discrete points of information within paragraph 3, which must be combined in order to respond correctly ('false'). 'Nineteenth century studies of the nature of genius' is contained within sentence 1 of the third paragraph, while the 'uniqueness of a person's upbringing' is discussed in lines 10-15 of paragraph 3. However, an accurate verbalisation of a correct response is unavailable to state conclusively the level of processing required for item 19.

Participant 2 also responded incorrectly to items 22-24 (inclusive). Explanations of her responses can be contrasted with correct responses to the same items by participant 5 to identify the necessary levels of processing for each item. Participant 2 initially selected the correct response to item 22 ('true'). Ultimately, however, she changed her answer to 'false'. The stem contains a comparative structure and therefore she identifies the key words in the question stem that occur twice ('the skills of') (coded 'syntactic parsing'). Although she declares uncertainty, she demonstrates an ability to 'establish propositional meaning at the clause level' when she correctly cites and explains the relative clause in sentence 1 of paragraph 5, although at this moment in the test, she cannot remember more, so circles 22 for further investigation:

> *"I circled item 22, and I underlined the key word 'skills' and I was sure about where I got this part, the key words from, paragraph 5. It's in the bottom part, I know the… general area this 'skills' I can retrieve from, and I circled 22 because I'm not sure about the answer. I think it's 'false', because I kind of got the idea from the text that people… the prodigies actually are similar to*

> *ordinary individuals; in essence they are similar, but they may show something different from ordinary people because of something; I'm not sure whether it's skills, or whether their minds, their brains are different. I'm not sure of that, but I thought it was 'false'. It's false"* [Participant 2, test 1, item 22].

Upon returning to item 22, she proceeds to paragraph 5. Line 3 had informed her of the similarity between genius' minds and those of ordinary people. However, once she consults this paragraph, she changes her response from 'true' (correct) to 'false' (incorrect). She states that initially, her choice was based on the congruence of meaning between the stem of item 22 and line 3 in paragraph 5, lines 2-3 ("…the manifestation of skills or abilities which are similar to, but so much superior to, our own"). Instead, the participant focuses attention on the subsequent sentence ("but that their minds are not different from our own is demonstrated by…"). She explains that her understanding of this sentence led her to alter her response; *"they are saying the skills are different, but their minds are not different"* [Participant 2, test 1, item 22]. The difference between the concepts of 'mind' and 'skills' led her to alter her response, evidence of 'establishing propositional meaning at the clause level'. She underlined 'minds' and 'not different' in the text to support her point.

In contrast, participant 5 did not appear to have any difficulty with item 22. She spent approximately 30 seconds reading paragraphs 4 and 5, before underlining part of paragraph 5, identifying a clause as relevant to item 22 ('which are similar to, but much superior to our own'). She does not highlight the subject or object in either the sentence or item, suggesting that she is able to hold this in her working memory quite easily (pronominal reference 'our own' relates to the head noun in item 22; 'ordinary individuals'). Highlighting the relative pronoun ('which') is evidence that she has established propositional meaning at the sentence level rather than just the clause level:

> *"the skills in a sense are the same as the skills of prodigies, because 'they are similar to, but so much superior to our own'. 'Our own' means the ordinary people, right?"* [Participant 5, test 2, item 22].

The participant spent no further time addressing this item. Establishing propositional meaning of the stem and relating it to a sentence paragraph 5 (rather than paragraph 3) was sufficient to answer this item type correctly.

Participant 5 reported difficulty with item 23, although she responded correctly, despite admitting in the interview that she was unable to fully comprehend it:

> *"Actually, I don't really understand this sentence [item 23]. Because it's too complicated… I was just thinking it ['lessen their significance'] can be the key point, like where they will ask for the answer. But, still, at that moment, I'm not really understanding the question."* [Participant 5, test 2, item 23].

She underlines 'lessen their significance' and then immediately returns to the text without having a clear idea of the meaning of the item, but with a conscious strategy of attempting to identify a relevant portion of the text to parse which will provide contextual meaning for item 23. She spends approximately 30 seconds reading paragraphs 4 and 5 carefully. She splits the stem of item 23 into constituent parts as a search strategy**.** She then returns to paragraph 5 and underlines the main clause of the final sentence ('minimise the supremacy of their achievements') and answers 'true' for item 23. She draws a parallel between this clause and the verb phrase in item 23 ('fails to lessen their significance'). She states that these two phrases are 'quite similar in meaning'. She then summarises the preceding information:

> *"I think it's quite similar to the meaning of 'lessen their significance'. Because it made the example for this kind of discoveries of knowledge is already being widely accepted. But it still doesn't mean they are not geniuses, they are, I marked here [between 'accepted' and 'and'] so I understand… because before I was like, I thought it should be here, these are two parallel sentences, so I was like, 'what's the meaning'? But after that, I figured out, OK, there should be like here, these are the part for which, and these are the main verbs, yep"* [Participant 5, test 2, item 23].

Despite the cognitive difficulty associated with this item, the relevant information was locally based in paragraph 5. This was therefore coded 'establishing propositional meaning at the sentence level'.

Participant 2 also has difficulty with item 23. She hesitates for about 10 seconds before selecting 'not given' for the item. She does not recognise the statement, meaning she is unable to reconcile the content with her existing knowledge of the text. The relevant part of paragraph 5 is written with explicit examples (Einstein, Kepler and Paul Klee). She displays no evidence of effectively parsing this statement, but can recognise that the words used do not appear in the text, suggesting that she is relying on a lexical strategy [lexical access], which is not sufficient to answer this item correctly.

Participant 2 responds incorrectly to item 24 ('false'; correct response 'not given'). Item 24 is also a complex sentence with a modified object noun phrase and subordinate clause linked to the main clause via a logical connective:

> ***24.*** *Giftedness and genius deserve proper scientific research into their true nature so that all talent may be retained for the human race.*

Participant 2 selects 'false' for item 24 (correct response; 'not given'). She states that this is "*kind of a guess*" [Participant 2, test 1, item 24]. Nonetheless, she underlines key words, suggesting she finds this item more accessible than the previous item [syntactic parsing]. She moves to paragraph 5 to reconsider item 24 after engaging with items 25 and 26. She identifies part of this paragraph and underlines a key word 'emulate' (line 4) and writes '24' next to it. She identifies this portion of the text ("*we try to emulate*…") as corresponding to researchers "try[ing] to… respond to giftedness and genius", although cannot state exactly why she responded false. This item does not have a clear textual reference. As the correct response is 'not given' participants may need to cite the item stem and explain the idea that is not located in the text. In this case, the word 'deserve' is the most significant clue. Participant 5 spent two minutes accessing portions of the text. She initially attempted a lexical access strategy then reads around the identified part of the text. She answers items

24 and 25 together, although in the interview addresses only item 25, not explaining her response to item 24.

Participants 2 and 5 both respond to items 25 and 26 correctly ('true' and 'not given' respectively), although participant 2 initially selected 'false' for item 25. This option was initially chosen in her initial responses to the items. She selects 'false' on the basis of the key words 'pay a high price' (which cohere with paragraph 6). She clearly remembers a part of the text which includes the key word 'price' but does not remember the proposition in which this key word is located. Therefore, this verbalisation is coded as 'lexical access'**.** She states that differentiating between 'false' and 'not given' at this point is difficult as she cannot remember the detailed information regarding the main point in the item stem. The participant returns to paragraph 6. She has the key words that she underlined form item 25 in her working memory ("pay a high price") and she identifies this in paragraph 6 (lexical access). She identifies the corresponding sentence in the text and underlines it in full ("recognise the price they may have paid in terms of perseverance, single-mindedness, dedication"). She explains that this is correlated with the wording of item 25. This is coded as 'establishing propositional meaning at the sentence level', as the participant identified information from multiple clauses as being synonymous with item 25. She changes item 25 from 'false' to 'true'. Item 25 is answered correctly. She then writes 'P6' next to item 25 for further reference if necessary.

Participant 5 also adopts a lexical access strategy ("I didn't find any exact information in the sentence matched to the option… I find the 'price', so I think it's related to the option [item] 25"), then reads around the identified part of the text. She does not find a synonym for 'high price' (item 25), but does recognise that the list of requirements in paragraph 5, lines 7-8 is a significant burden, so she answered 'true' for item 25. That she recognises the pronoun 'they' refers to 'geniuses' suggests she combined information from more than one clause, so this stands as evidence of 'establishing propositional meaning at the sentence level'.

In relation to item 26, participant 2 states that she did not find any evidence for item 26, so responds with 'not given'. She correctly recognises that the phrase 'high personal cost' does not appear in the text ('lexical access'). She also displays some test-wise behaviour, on the

basis of her understanding of the test format, and the number of responses in each category ("I want two NGs", participant 2, test 1, item 26). Participant 5 recognises that item 26 is an attitude statement. She scans the final two paragraphs for such a statement. She is able to summarise paragraphs 6 and 7, providing evidence that she has established a mental model of the final two paragraphs after engaging with the text to address these items. This is the only example of participants using higher level processing to respond to this item type. She confirms that no statement is given by the author regarding whether becoming a genius is worth the high personal cost:

> *"I don't think they really made a conclusion that it is worth to pay such a high price to be a genius. It just says here we teach or educate our kids in kind of like encourage them to be a genius, but we need also not to forget that you had to make a lot of effort to be a genius, but the author didn't really say it is worth [it] or not. And in the last paragraph, I as just making sure of that"* [Participant 5, test 2, item 26].

A high proportion of 'lexical access' (fourteen instances) is consistent with strategic management of identifying key words in this item type. Evidence suggests that other frequent levels of processing for correct responses to this item type are 'establishing propositional meaning at the clause' and 'sentence' levels, with seven and eight instances emerging of each respectively.

4.5.1.4. Type 3: Identifying writer's views/claims (3-IDw)

Strategy code	Strategy label	Freq. rate
11	Careful local reading (text)	**2.25**
27	Compares question stem/option to a portion of the text	**2.25**
6	Marks/notes key noun phrase(s) in the questions stem or options	**1.00**
28	Checks/confirms/considers option choice after reading portion of text	**1.00**
19	Searches for key word/phrase (text)	**0.75**
18	Returns to the question for clarification: rereads question and/or options	**0.50**
24	Identifies paraphrase within text or between text and item stem	**0.50**

Table 4.23. Most commonly-used strategies for IELTS Identifying writer's views/claims items

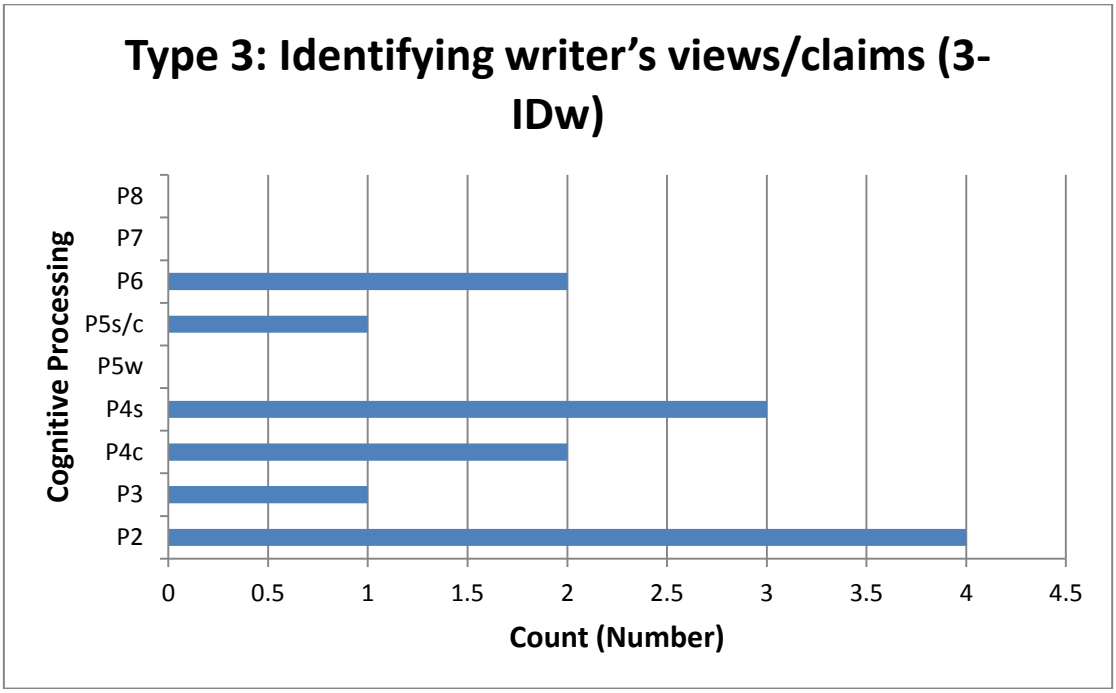


Figure 4.6. Frequency of cognitive processes for IELTS Identifying information (True/False/Not given) items

	P2	P3	P4c	P4s	P5w	P5s/c	P6	P7	P8
Total Processes	**4**	**1**	**2**	**3**	**0**	**1**	**2**	**0**	**0**
Processing ratio	**1.00**	**0.25**	**0.50**	**0.75**	**0.00**	**0.25**	**0.50**	**0.00**	**0.00**

Table 4.24. Most frequently-identified cognitive processes for IELTS Identifying information (Yes/No/Not given) items

Sample item

Questions 37—40

Do the following statements agree with the views of the writer in Reading Passage 3?
In boxes 37-40 on your answer sheet, write

YES	*if the statement agrees with the views of the writer*
NO	*if the statement contradicts the views of the writer*
NOT GIVEN	*if it is impossible to say what the writer thinks about this*

37 The wear and tear theory applies to both artificial objects and biological systems.
_____ **[NO]**

Textual reference: "A further argument against the simple wear and tear theory is the observation that the time within which organisms' age lies between a few days (even a few hours for unicellular organisms) and several thousand years, as with mammoth trees"

'Identifying the writer's views/claims' is a similar item type to 'identifying information'. Participants are presented with a sentence, which is a restatement of part of the text and are asked whether this statement is an accurate reflection of the content ('yes'), contradicts information in the text ('no'), or is not stated in the text ('not given'). However, in this item type, the statements presented to test takers reflect claims made by the author or by named individuals within the text. As a result, this item type is more likely to be used with discursive texts which present an argument, or opposing opinions or claims. Four examples of this item type were included in the test instrument (text 3, items 37-40). These items were completed by participants 3 and 6. Both participants answered all three items correctly.

Only four strategies receiving ratings of 'very high' or 'high'. These strategies suggest a common approach of identifying key nouns in item stems and identifying relevant parts of the text containing those words. The most commonly observed behaviour was close, careful reading of the text. Participants moved between the text and items frequently, like the previous item type. However, for 'identifying the writer's views/claims', participants displayed a greater tendency to use higher level processing. As this item type constituted the final three items of the IELTS test, the participants may have been more likely to report their understanding of the text as part of the explanations of how they completed this item type. Overall, establishing propositional meaning at the clause and sentence levels, and establishing a mental model of part of the text were the most frequently used codes.

For item 37, participant 3 reads paragraphs C and D for approximately thirty seconds before writing 'no'. For this item, she focused on the head noun phrase ('wear and tear theory'), which she was able to match to paragraph D (lexical access). She identifies the noun phrase/sentence subject containing 'the wear and tear theory' and notices that the author seems to criticise the theory. She re-reads the question stem, noting the second half (coded 'establishing propositional meaning at the clause level'). She writes 'no' on the basis that the first half of the sentence in the text she identifies suggests that the writer is *dismissing* the

theory rather than using it in her argument. At this point, there is no evidence of her having parsed the complete sentence. For this reason, she double-checks her response immediately, returning to the text at 2.27, spending approximately 15 seconds checking:

> *"For this question (Q37), I was looking at this, the 'wear and tear theory' and then I remember, somewhere in the passage, it talks about the 'wear and tear theory', so I was really just scanning to pinpoint the location of this concept, and then I found it in paragraph D, so I read the sentence 'a further argument against the wear and tear theory', so apparently it sounds like the author is criticising the simple wear and tear theory, so I read the sentence [item 37] again, it said 'applies to both' and then I realised this should be wrong, because actually the writer's view is against the wear and tear theory, it's not applying the wear and tear theory in his argument, so I wrote 'No' there"* [Participant 3, test 1, item 37].

The participant clearly used her knowledge of the text to be able to respond to the item within thirty seconds. She used a paragraph with detailed information about the 'wear and tear theory' and was able to use a single example of text that contradicts it to identify the correct response. Participant 6 referred to a different part of the text (paragraph A). He also uses the strategy of identifying the principle noun phrases to lead him to the relevant portions of the text ('wear and tear theory') (lexical access), in this case paragraph A. He identifies a relevant sentence and compares it to another in the same paragraph as essentially a paraphrase, successfully inferring at the clause level the meaning of the 'wear and tear' theory rather than parsing a definition:

> *"It [item 37] mentioned 'wear and tear' theory… The sentence in this statement, 'we think of artificially produced technical objects, products which are subject to natural wear and tear during use' [paragraph A], this is… 'Wear and tear' and 'loss of function of technical objects' seems really similar"* [Participant 6, test 2, item 37].

However, participant 6 does not display evidence of further engagement with the text in relation to item 37, providing no further evidence of how he arrived at the correct response. Paragraph A outlines the problem that is restated in item 37, but does not answer the issue. Therefore, the participant displays evidence of high level processing, although not to directly answer the item.

Participant 3 responds to item 38 by selecting 'yes'. Item 38 is a complex sentence, placing a large cognitive load on the participants. She then went to paragraph F, believing it to contain the relevant portion of the text. She found no explicit textual reference, although her understanding of the text at this point led her to answering 'yes'. She returned to this question after she had completed the others, she realised that this paragraph should not contain the answers to more than one item (a test-wise strategy), so she would have to conduct a new search. For this reason, she moves to paragraph B and identifies the relevant sentence. She recognises the synonymy of these two complex sentences and retains her answer. As a single sentence provided sufficient evidence to answer correctly, this was coded 'establishing propositional meaning at the sentence level':

> *"[For] question 38, my instinctive answer is 'Yes', because I remember in one paragraph, it talked about how energy consumption, like energy preservation can help prolong your life, right, then I thought 'yes, you can grow older, but you may not age very much simply because your way of living and the pattern of energy consumption. So that's why I went to paragraph F and that was the paragraph about the energy consumption, so after I read that I can't be sure, but I think after that I put 'Yes' there. But, interestingly, after I finish question 40 is something from that paragraph as well and that's got me thinking, so question 38 shouldn't be from the same paragraph, and that's when I read… I think I went back to paragraph B and read again, and then I found the exact sentence where it says 'as long as a biological system has the ability to renew itself it could actually become older without ageing'. So it's almost like the same sentence. That's when I confirmed that from both paragraphs, B and F. Reading paragraph F was a way of confirmation of my previous assumption about this, from memory"* [Participant 3, test 1, item 38].

Answering item 38, participant 6 moves to paragraph B. In answering items 27-32, the participant highlighted sentence 5 in paragraph B, which he cites as evidence in returning to paragraph B for item 38. He correctly summarises this sentence. The first preposition phrase of the item ('in principle') is parsed in relation to the first clause in sentence 5 in paragraph B ('as long as the biological system has the ability to renew itself…'), which the participant links confidently. This is evidence of establishing propositional meaning at the sentence level:

> *"It's talking about ageing and the biological system. I think it has been mentioned in paragraph B, so I just look back to paragraph B. Because the sentence 'at least as long as a biological system has the ability to renew itself, it could actually become older without ageing'. That's what the question is asking [about]. Because as long as a system has the ability to renew itself, the precondition, it can become older without ageing, so it's possible"* [Participant 6, test 2, item 38].

Participant 3 progressed to item 39 and reads the stem before returning to paragraph B. About 40 seconds later, she selects 'not given' for item 39. While reading paragraph B for item 38, she states that she also kept item 39 in her working memory;

> *"When I was reading paragraph B at that moment, the moment [3.50], I was actually thinking about 39, because here it's talking about replacement of 'genetic material'",* [Participant 3, test 1, item 39].

She was able to draw a parallel between the item stem and the content of paragraph B, although does not cite any specific sentences or textual references, so this can only be coded as 'lexical access'. But then she went into this paragraph in more depth ("I spent quite some time reading paragraph B again"). She was searching for any information related to direct statements in which proportions of replacement of genetic material was mentioned (the stem states 'about 90 per cent'), but she found none, so selected 'not given':

> *"I spent quite some time reading paragraph B again. I knew that in this paragraph, it talked about how new material replaced old material, but I wanted to make sure that actually whether it talked about proportions of replacement and that's when I read again, and noticed it was not mentioned in any place in this paragraph, so I went back and I wrote 'Not given'"* [Participant 3, test 1, item 39].

Participant 6 also correctly answers item 39 ('not given') as he focuses on the proportion cited in the item ('about 90 per cent'). As there is only 30 seconds between answering items 38 and 39, the participant must use his working memory of the text to state that there is no such statistic mentioned in the text, although such knowledge does not constitute a mental text-level understanding. To correctly answer this item, establishing propositional meaning at the clause level to parse the item is the minimum level of processing required:

> *"I don't think it mentions [in] the text, 'about 90% of the human body is replaced as new', there is no statistic[al] number for this. So I just write 'not given'. I think it's not given in the text"* [Participant 6, test 2, item 39].

For item 40, participant 3 reads the stem and immediately answers 'yes'. When attempting to answer item 38, she had already read paragraph F extensively and had a clear understanding of the paragraph (coded 'building a mental model'), so was able to select her answer without further reference to the text:

> *"This [one], 'conserving energy may help to extend a human's life', I wrote down 'Yes' immediately because when I was doing number 38, I already read paragraph F again, so I and also when I was doing the first part, I developed a very clear idea about that paragraph, so that's why here, I wrote down 'Yes' immediately, but as I talked about before, when I wrote down 'Yes' for this question, I realised 'oh, maybe somewhere else talks about question 38, so now that's when I went back to paragraph B again"* [Participant 3, test 1, item 40].

Participant 6 responded to item 40 before item 39 (about 15 seconds after item 38). He selected 'yes' on the basis of linking 'conserving energy' to paragraph F. He states directly that he remembers the content of paragraph F, that you can "extend your life" (the participant directly links this item to item 32 (paragraph G heading), and heading 4 ('prolonging your life'), calling into question the independence of items 32 and 40). He is able to answer the question without returning to the text, suggesting that his statement that he has built a mental model of paragraph F is accurate:

> *"It's about 'conserving energy', so it's mentioned in paragraph F. I remember the content, if you stay calm, you can extend your life, so energy can help you to extend your life, so I put 'yes' here"* [Participant 6, test 2, item 40].

4.5.1.5. Type 4: Matching information (4-MI)

Strategy code	Strategy label	Freq. rate
6	Marks/notes key noun phrase(s) in the questions stem or options	**2.40**
33	Writes note/labels part of text/item	**2.00**
18	Returns to the question for clarification: rereads question and/or options	**1.80**
7	Marks/notes key verb phrase(s) in the question stem or options	**1.80**
8	Marks/notes key adjective phrase(s) in the question stem or options	**1.60**
12	Marks/notes key noun phrase in the text during careful reading	**1.60**
11	Careful local reading (text)	**1.40**
27	Compares question stem/option to a portion of the text	**1.40**
24	Identifies paraphrase within text or between text and item stem	**1.00**
19	Searches for key word/phrase (text)	**0.80**
26	Eliminates option (s) (no information found)	**0.80**
13	Marks/notes key verb phrase in the text during careful reading	**0.80**
14	Marks/notes key adjective phrase in the text during careful reading	**0.80**
30	Hesitates while answering to reconsider choice	**0.60**
5	Reads question stem(s) and/or option(s) carefully	**0.60**
28	Checks/confirms/considers option choice after reading portion of text	**0.40**
2	Identifies the purpose of the question	**0.40**
9	Marks/notes key prepositional phrase(s) in the question stem or options	**0.40**
25	Eliminates option(s) (contradicts key noun/verb phrase/nonsensical response)	**0.40**

29	Guessing	**0.40**

Table 4.25. Most commonly-used strategies for IELTS matching information items

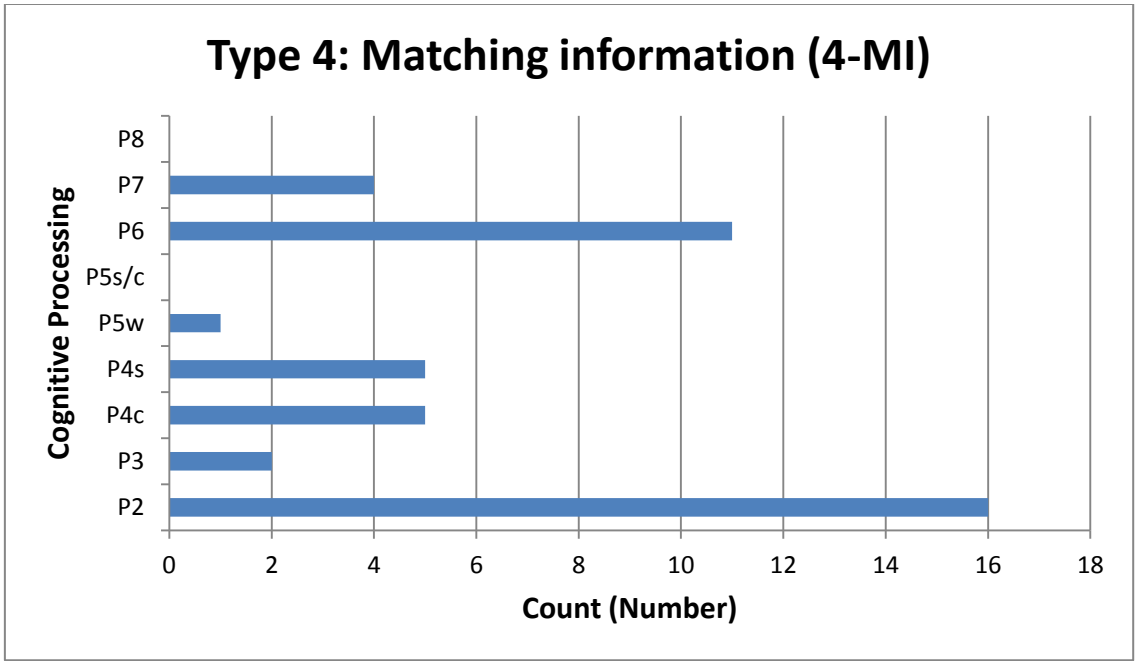


Figure 4.7. Frequency of cognitive processes for IELTS Identifying information (True/False/Not given) items

	P2	P3	P4c	P4s	P5w	P5s/c	P6	P7	P8
Total Processes	**14**	**10**	**8**	**9**	**0**	**1**	**2**	**2**	**0**
Processing Ratio	**2.80**	**2.00**	**1.60**	**1.80**	**0.00**	**0.20**	**0.40**	**0.40**	**0.00**

Table 4.26. Most frequently-identified cognitive processes for IELTS matching information items

Sample item

Questions 14—18
*Choose **FIVE** letters, **A—K**.*
Write the correct letters in boxes 14—18 on your answer sheet.
***NB** Your answers may be given in any order.*
Below are listed some popular beliefs about genius and giftedness.
Which **FIVE** of these beliefs are reported by the writer of the text?

A Truly gifted people are talented in all areas.
B The talents of geniuses are soon exhausted.
C Gifted people should use their gifts.
D A genius appears once in every generation.
E Genius can be easily destroyed by discouragement.
F Genius is inherited.

G	Gifted people are very hard to live with.
H	People never appreciate true genius.
I	Geniuses are natural leaders.
J	Gifted people develop their greatness through difficulties.
K	Genius will always reveal itself.

Items 14-18 are 'matching information' items which require test takers to identify which of a series of sentences reflect accurate statements that are made in the text. In this test, participants are asked to match common beliefs to those that are cited in the text and eliminate those that are not. Participants must select five out of eleven options. The content of the sentences will reflect the genre of the text. The number of items associated will depend on the number of options that the test takers are expected to identify. In this test, this item type accounts for five items. A range of strategies are employed for this item type, although only two were rated as 'very high'. Additionally, seven strategies were rated as 'high'. Most noticeably, these involved marking key words (verbs, adjectives and nouns) in both the text and options, moving between the text and options for clarification and to identify paraphrases and for comparison purposes. The highly varying strategic actions suggests multiple approaches were adopted to answer these items.

The majority of the cognitive processes observed for this item type were lower level processes. Lexical access and syntactic parsing were rated 'very high', with establishing propositional meaning at the clause and sentence levels rated as 'high'. However, some evidence of higher level processing in relation to the item type also emerged. Forming a mental model and establishing text level representation were rated 'medium', indicating that participants' knowledge of the text assisted them in responding. Examples of this item type occurred only in relation to text 2 completed by participants 2 and 5. Both participants had mixed success with this item type, each recording three correct responses from 5. However, a correct response for each of the keys was obtained. Combined with analysis of incorrect responses, this provided a lucid picture of relevant item completion processes. Item analysis in this section follows participant responses. Participants tended to keep several options in their working memory and read parts of the text for multiple pieces of information simultaneously.

Participant 2 begins by reading paragraph 1 carefully. She has a specific strategy for reading carefully (placing a line after each sentence or clause] to break it up into more manageable chunks. Verbalisations at this stage refer to the strategy, not linking what she is reading to any of the question options:

> *"I'm underlining words that I think can help me later with the question I'm going to answer… 'genius [and] 'giftedness'; that's the topic of this reading passage and the requirements of this item. I need to underline five of the beliefs I need to find… a sentence has a meaning that makes sense to me. This sentence is very long, and I try to find the lexical items… the head noun, the head of the sentence [attributes, characteristics]"* [Participant 2, test 1, items 14-18].

After having read the first three paragraphs, the participant returns to the question options and marks three of the options. Option A is eliminated (correct), option B is selected (correct) and a mark is drawn next to option D to show that she is still considering it. She references paragraph 2 in support of selecting option B, although does not supply a specific reference. Therefore, the highest level of processing that may be assigned to these observations and verbalisation is syntactic parsing. She then considers option E, marking it with a hyphen. She states that there is no affirmative evidence of this statement in the first two paragraphs, which is correct. Moving from paragraph 3 to option E and considering this option is a process that lasts approximately 8 seconds. This is insufficient time to reconsider the content of the previous two paragraphs, indicating that she has successfully built a mental model of these paragraphs and is establishing whether option E coheres with the mental representation that she is storing in her working memory of the first two paragraphs:

> *"I underlined the words 'destroyed' and 'discouragement' [option E] and I tried to look at, because this is the part as far as I went at that moment, so I guess the 'difficulties and adjustment' if they are related. Because I guess I was sure, there is nothing related in the second paragraph and the first paragraph that*

is related to this question, to this item or this statement" [Participant 2, test 1, items 14-18].

Further evidence of the participant employing a mental model is evident in her action at 4.45, in which she selects option F (correct). She answered correctly without referring back to the text. Option F coheres with a proposition established in the mental model of paragraph 1. Khalifa and Weir (2009: 52) note that returning to check (monitoring) information relating to a mental model occurs when incoming information contradicts the established proposition. As option F coheres with the existing representation, the participant selects this option with confidence. Additionally, there is evidence of establishing propositional meaning and inferencing at the sentence level, as the participant links the notion of heritability in option F to the final sentence in paragraph 1 and references 'genes or genetics' as 'the source of exceptional abilities' (*"I answered from the first paragraph, the 'genes', 'genetics', I think this is right, correct"*, participant 2, test 1, items 14-18).

The participant is unable to establish the accuracy of options C and G from her knowledge of the text, so returns to paragraph 3. She underlines two portions (lexical access). She stated that she was attempting to find parts of the text relevant to options C and G. She then continues reading paragraph 3 and moves to paragraph 4. She eliminates options I (correct) and J (incorrect) on the basis that she did not find any relevant textual references paragraphs 1-5. She was confident in eliminating option I, indicating successfully established propositional knowledge at the clause level of option I**:**

"I guess that is the statement [paragraph 4, line 3] that I need to read, I guess I had read the previous ones, the ones above the sentence; it's in the order I covered these statements… I didn't find the answer in paragraph 4 [or] paragraph 5… [Participant 2, test 1, items 14-18].

She eliminates option J (incorrectly) by referring to paragraph 3 ('difficulties of adjustment') but does not explain how this textual reference contradicts the information in option J. There is no evidence of processing beyond lexical access at this point. She then selects

option K on the basis that she has gone through all of the options and has not yet selected 5, so selects option K as by a process of elimination it is likely to be right. She does not refer back to the text, and as option K does not refer to a specific portion of the text, if she had noticed the lack of coherence between this item and her existing mental model, Khalifa and Weir (2009: 52) suggest that she should go back to the text to monitor the macro and micro propositions of her mental model. Instead, once she has responded, she goes on to read the remaining paragraphs. As she admits that she guessed at this point, this verbalisation and associated actions were coded as test-wiseness.

The participant skimmed paragraph 6 in order to gain a general understanding before reading in depth. She identifies the topic sentence and repeats this pattern for paragraph 7. She returns to paragraph 2, where most of the textual references for the answer keys are located. She identifies line 1 and links it with option H ('envy' and 'never appreciate'; correct text reference), suggesting that she has established propositional knowledge at the sentence level. At this point, she declares that paragraph 2 is the most important part of the text for her purposes, indicating that she has sufficient knowledge of the text to navigate effectively. Although she cites key words (lexical access) as evidence for why she focuses on paragraph 2, there is sufficient evidence from previous verbalisations and observations that she has engaged with the text at the propositional clause and sentence level and has thus created mental representations of each of the paragraphs:

> *"After I ticked the item key, I guess I finished all of the statements, and you can see I spent several minutes re-reading the last paragraphs, I tried to get a general idea and try to make sure, make certain that there isn't any evidence inside the three paragraphs. I was thinking [that] I found this paragraph [2] might be the most important paragraph I need to read and to re-read again, so I'm kind of doing intensive reading. A lot of the words I have read from the items of the test, there are a lot* of *key words from the item sentences, I guess, I read from paragraph 2"* [Participant 2, test 1, items 14-18].

In reviewing her choices, she returns to paragraph 2 and identifies a clause that she states is evidence backing up her initial selection of option K ('genius will always reveal itself'), although this is a contradictory statement, so cannot be used as evidence of establishing propositional meaning. Instead it is coded as syntactic parsing on the basis that she is able to parse the clause and recognise its structure, if not the propositional content. She then eliminates option C, stating that she can find no evidence for it in paragraph 2 (although this option is correct and the relevant textual reference is located in paragraph 2, line 9).

Participant 5 begins by familiarising herself with the text rather than the instructions or the items. She quickly reads the title and paragraph 1, but does not highlight or underline anything. She explains that her exam experience taught her to try and understand what the text was about prior to reading the items. She briefly moved to paragraph 2 and highlighted the word 'concept' (lexical access) ("*this paragraph, I guess, will probably be related to 'concepts'*", participant 5, test 2, items 14-18), but does not explore further. She states that paragraph 1 was easy to access (with the exception of the term 'prodigies') but paragraph 2 contained more complex ideas which were difficult to parse. She moves to the question options and engages with them at the word level. How she interacts with each item is illuminating in terms of her cognitive processing. The words that she highlights provide evidence of establishing propositional meaning at the clause level. She only highlights the noun phrase first in options A and I. In options B, E, F and G, she highlights key adjective phrases, and in options C, D, H and K, highlights verb phrases. As the text relates to genius, the majority of the noun phrases in the options relate to genius or giftedness. Options A and I relate giftedness to additional concepts, such as whether talent extends to all areas of genius' lives (A), and leadership (I). Adjective phrases describe individual beliefs, so matches the requirements of the question. Verb phrases, as the 'glue' that binds the clauses together, are highlighted when they display particular attitudes or sentiments towards genius ('should use' and 'never appreciate'). Adverbial phrases provide impetus to look for adverbs in the text and compare the meaning of the clauses in which they are located to those in the options:

> *"I just circled points that I think [contain the] main meaning. The key words. It just helped me to think. I would circle the part which I think, like*

> *'in all' [option A] of these kind of words, are controversial, because 'in all' maybe it's what they really do, maybe in the article they only say some of them… they didn't say 'all'. Then it could be like a place where they make a trick. I circled the part [which] is most important, and underline the words like 'all', or 'talented', then the other words and this here, like 'once'. I think 'once' should be reconsidered, but 'once [what] in every generation'"* [Participant 5, test 2, items 14-18].

She then examines the question stem ('which FIVE…') (lexical access) before returning to the text and reading paragraph 2. She underlines sentence 2 in paragraph 2 and directly links this to option H, declaring them to be synonymous. Sentence 2 is a short, compound sentence ('we envy the gifted and mistrust them'). This is a simple, but nonetheless multi-clause sentence. The participant demonstrates ability to link the two main verbs to 'never appreciate' in option H (correct) (*"Because I think this sentence means similar to sentence H. We never appreciate the genius, so I marked, I wrote 'H' here",* participant 5, test 2, items 14-18*)*. Therefore, this is an example of establishing propositional meaning at the sentence level. Shen then continues reading paragraph 2, expending a considerable amount of time parsing and understanding paragraph 2 (at least 1 minute, 17 seconds) before her next observable action, which is to underline a portion of the line 4, paragraph 2. At 4.37, she writes 'A' and 'B' in lines 3 and 4 respectively, indicating that when she returned to the options at 4.16, she held information from multiple clauses in her working memory.

The majority of the popular beliefs for questions 14-18 come from paragraph 2. The phrase 'popular beliefs' in the question stem is paraphrased in line 3 of paragraph 2, although there is no evidence that the participant used this verbal cue to initiate a search in this portion of the text. The parts of the text that she labelled for options A and B come from a single compound sentence. For option B, she highlighted a compound sentence (establishing propositional meaning at the sentence level) and links key words in option B to the text ("*I go on to reading, and here, I highlight this part, it's because I think this is quite similar to this one [option B], ''are soon exhausted'… 'too brightly, too soon and burn out'",* participant 5, test 2, items 14-18). Paragraph 2 contains a sentence which is a list of co-ordinated clauses that are subordinate to the main clause 'it is popularly believed…'. There is evidence that

she linked this co-ordinate clause to the main clause in the sentence ("It says 'beliefs', so it basically means popular beliefs"). Evidence suggests she failed to link the subordinate clause used for option A to the main clause the follows it, as she highlights only the subordinate ('if people are talented in one area…') clause, but not the main clause, leading her to incorrectly select option A. Had she continued parsing this sentence, she would likely have disregarded this option ('…they must be defective in another'). Therefore, for option A, she only established propositional meaning at the clause level.

The participant continues reading the remainder of paragraph 2. She is not satisfied that she has identified all of the relevant beliefs, so she continues in to paragraph 3, underlining a noun phrase ('two of the most significant') as the key point that the paragraph addresses. She spends some time reflecting on the meaning of the first few lines of paragraph 3 to determine whether this portion is required to complete questions 14-18. She admits at this stage that she is struggling to establish propositional meaning of the sentences in paragraph. Therefore, the highest code that was attributed to this part of the test is syntactic parsing, on the basis that the noun phrase she highlights indicates the purpose of the paragraph. She then returns to the options and underlines the letters that she has selected at this stage of the test:

> *"Because I'm not very sure that I get it. I think one of my problems is like every single word, I understand it, but when they [are] put together, I was like, I'm not very good at grammar. When I wrote down this [the five selected options], I was quite sure that these are the answers.* [Participant 5, test 2, items 14-18].

Items 14-18 were centred on one paragraph (paragraph 2), relating to the list of beliefs that the author cites. Success depends on parsing the first part of the sentence and relating that to the task instructions. Based on item design and participant responses, this item type targets lower-level processes, prioritising the need to construct propositional meaning at the clause and sentence levels. Occasionally, participants used their knowledge of the text to build their confidence in using one part of the text to address these questions while disregarding other parts. Concentrating five items in close proximity in a single paragraph

created challenges in test management as the participants were not anticipating that all of the answers would be situated in one paragraph.

4.5.1.6. Type 5: Matching headings (5-MH)

Strategy code	Strategy label	Freq. rate
11	Careful local reading (text)	**1.50**
18	Returns to the question for clarification: rereads question and/or options	**1.33**
27	Compares question stem/option to a portion of the text	**1.00**
28	Checks/confirms/considers option choice after reading portion of text	**0.83**
30	Hesitates while answering to reconsider choice	**0.67**
6	Marks/notes key noun phrase(s) in the questions stem or options	**0.33**
12	Marks/notes key noun phrase in the text during careful reading	**0.33**
24	Identifies paraphrase within text or between text and item stem	**0.33**
19	Searches for key word/phrase (text)	**0.33**
14	Marks/notes key adjective phrase in the text during careful reading	**0.33**
5	Reads question stem(s) and/or option(s) carefully	**0.33**
2	Identifies the purpose of the question	**0.33**
1	Reads question(s) before proceeding to the text	**0.33**
22	Identifies content-based parallel between paragraphs	**0.33**
20	Skimming part of the text for general understanding (expeditious reading)	**0.33**
23	Identifies lexical parallel between parts of the text (matches words across paragraphs or heading)	**0.33**
32	Uses own topic knowledge to enhance understanding of text/questions	**0.33**
31	Checks/confirms response has met the parameters of the task	**0.33**

Table 4.27. Most commonly-used strategies for IELTS matching heading items

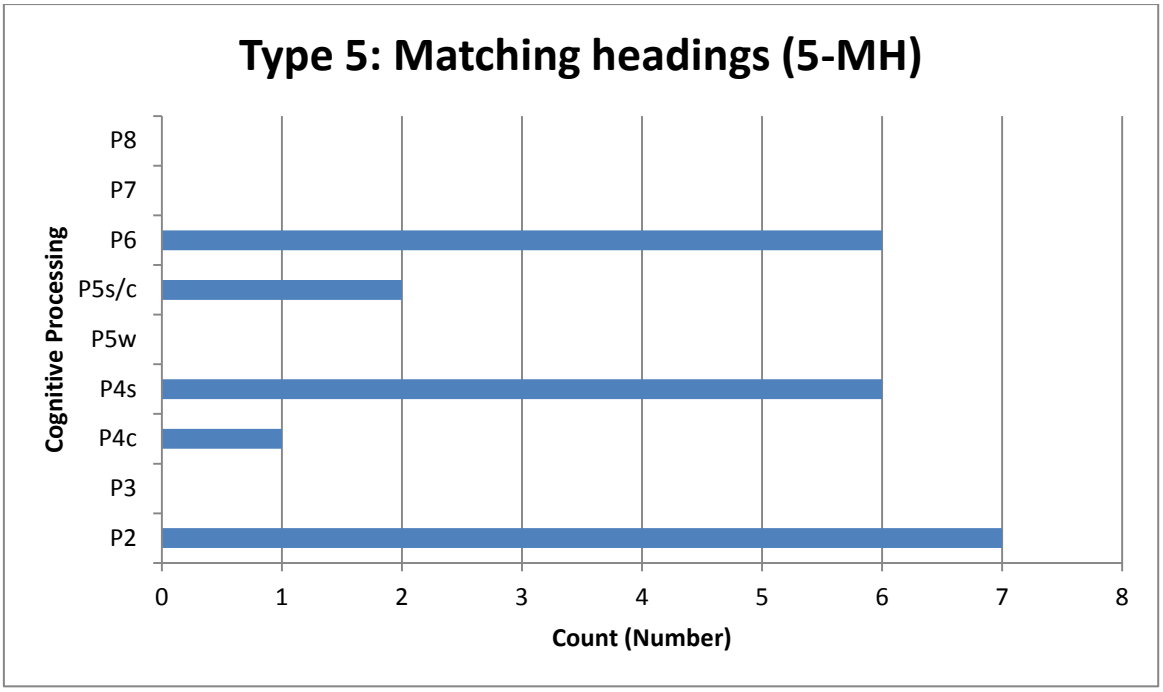


Figure 4.8. Frequency of cognitive processes for IELTS matching headings items

	P2	P3	P4c	P4s	P5w	P5s/c	P6	P7	P8
Total processes	**7**	**0**	**1**	**6**	**0**	**2**	**6**	**0**	**0**
Processing ratio	**1.17**	**0.00**	**0.17**	**1.00**	**0.00**	**0.33**	**1.00**	**0.00**	**0.00**

Table 4.28. Most frequently-identified cognitive processes for IELTS matching headings items

Sample items

Questions 27—32
*Reading Passage 3 has seven paragraphs, **A—G**.*
*Choose the correct heading for paragraphs **B—G** from the list of headings below.*
*Write the correct number, **i—x**, In boxes 27—32 on your answer sheet.*

List of Headings

i	The biological clock
ii	Why dying is beneficial
iii	The ageing process of men and women
iv	Prolonging your life
v	Limitations of life span
vi	Modes of development of different species
vii	A stable life span despite improvements
viii	Energy consumption
ix	Fundamental differences in ageing of objects and organisms

x	Repair of genetic material

'Matching heading' items present test takers with a series of noun phrases (only title 2 is a clause containing a verb), which they are required to link to the main purpose of each paragraph. Unlike 'matching information' items, given headings align to specific paragraphs, and participants are required to assign each paragraph a heading. The number of headings that need to be selected corresponds to the number of paragraphs in the text. Paragraphs are lettered to assist test takers. The test developers clearly intend that this item type should encourage test takers to engage with each paragraph carefully, in order to identify the main purpose of each. The main purpose may be included in the first sentence of each paragraph, although this is not universally true, requiring test takers to engage more closely to distinguish between information which is presented as a main argument and information which acts to support that argument.

'Matching heading' items are connected to text 3 ('how does the biological clock tick?'). This text contains seven paragraphs. Paragraph A is assigned a heading as an example. These items were completed by participants 3 and 6. Test takers are required to identify headings for the remaining six paragraphs, meaning there are six items. There are ten options. Participants' strategic actions were mixed; only three strategies recorded 'high' use and a further two 'medium' use. The remaining strategies cited recorded 'low' use. These items inspired close engagement with the text and participants moved frequently between the paragraphs and the headings to read, consider, select and eliminate options based on their engagement. Hesitation was also observed, indicating close engagement with the text as participants considered a heading in relation to a paragraph. Both participants performed well in this part of the test, recording 5/6 correct. Participant 3 answered item 27 incorrectly, and participant 6 answered item 28 incorrectly.

There is strong evidence from the engagement of the two participants that this item type tests a range of both higher and lower cognitive processes. The items are presented to the test takers before the text, as the test developers prompt the test takers to read the headings before proceeding to the text. To link these titles to paragraphs requires a general understanding of the topic of each paragraph and an ability to summarise the main ideas.

Participant 3 was able to describe metacognitive strategies at each stage of task completion. She stated that she attempted to form a mental model of each paragraph after reading and then try to match this model to one of the options in the list of headings. She also reports remediation at moments that she decided that the chosen strategy was not efficient, and critiques individual strategies that she was aware of but consciously rejected *("some people do key words, they underline key words and things, I was thinking about doing that, but then I thought that 'is that too simplistic?"*). Participant 3 did not display much observable interaction with the instrument. She did not underline, highlight, or trace her progress through the text. She only wrote on the sheet when she was satisfied with a response. Despite the paucity of visual stimuli in the interview, she is able to report how she engaged with the text at specific moments ('I was sort of seeing if I could read this very quickly, and identify the main ideas from the list'). As the list of headings largely contains noun phrases, they are easier to hold in the working memory than complex sentences which express ideas, rather than concepts.

At 2:01, she reports considering an option in relation to item 28 (paragraph C). She references the third title ('the ageing process of men and women') and links this to the words 'ageing and death' in paragraph C (lexical access); although she realises that this is an insufficient basis on which to choose this option. She then expands her understanding of the paragraph and states that the textual reference 'immortality would disturb this system indicates that dying is necessary, linking paragraph 3 to title 2 ('why dying is beneficial'), which she selects. Nowhere in the text does it state that dying is beneficial; the participant therefore infers this information:

> *"This is the paragraph where it talks about ageing and death, and um… I think this paragraph, it talks about how age and death makes possible the biological systems to maintain… balance, almost suggesting the idea that it's necessary for ageing and death to take place. So that's why I chose number 2"* [Participant 3, test 1, items 27-32].

As she progresses, she notes that in several paragraphs, the term 'life span' is mentioned (lexical access) and she also draws a lexical parallel between the item stem of title 1 and

paragraph D ('biological clock'). She notes at this point that she did not commit to any choices until she had read all of the text. As she moves through the text, she explains her understanding of paragraph F ('something about how animals consume energy and how that relates to their longevity'), evidence of her building a mental model as she has parsed the paragraph and identified which are the most important points needed to summarise the paragraph. There is no evidence that participant 3 formed an overall text-level representation; task requirements did require close engagement with the complete text. As paragraph A is included as an example item, this paragraph was dismissed from any consideration, ruling out her forming a text-level representation. She states that she has a clearer understanding of paragraphs E, F and G, but C and D she is less sure of. Her strategy at this point is to move between the paragraphs and the list of headings to narrow the search.

At 4.57, she selects option 3 for item 27 (incorrect). From the second half of paragraph B, she deduces that it is a discussion of the ageing process (agreeing with the first noun phrase in heading 3), but she concedes that men and women are not mentioned in the paragraph (the second noun phrase in the heading). She displays evidence of inferencing at the clause level by linking 'the destruction of old material and the formation of new material' to 'the ageing process':

> *"I feel like this paragraph is about 'process'; it talks about the 'destruction of old material and the formation of new material', so that's what makes me feel like this should be about the ageing process, but to be honest, it says men and women and I feel like it doesn't mention anything about men and women, so I did hesitate a little bit about this one, but I sort of went through that and felt like that's the closest answer, so that's why I chose this one. But now I'm thinking about it, I feel that maybe it's 'repair of genetic material'"* [participant 3, test 1, items 27-32].

The correct response to item 27 (paragraph B) is title 9 ('fundamental differences in ageing of objects and organisms') which requires recognition that sentences 1 and 2 of paragraph B

contrast with sentences 4-6, linked by transitional sentence 3. This suggests that answering this item correctly requires the formation of a mental model, which participant 3 displays no evidence of having done. This contrasts with the verbal record of participant 6, who states the following:

> *"I was trying to take [select] one for paragraph B. I think it's a comparison here… I found 'organism' in the second part, and in the first part, it talks about the object[s], they are talking about differences between those two things. It's this sentence; 'although the same law holds for living organism, the result of this law is not inexorable in the same way'. This sentence helps me to think it's a comparison between the second thing and the first thing. Because although the same law holds, it's not in the same way"* [participant 6, test 2, items 27-32].

After briefly re-familiarising herself with the headings, she selects option 9 for item 27. She identifies key words pertinent to the heading (organism, object) (lexical access), and clearly states that the paragraph discusses differences between these two concepts by citing the fourth sentence in the paragraph. The cited sentence includes only one of the key words, so he must have successfully parsed sentence two (the contrast is marked with the preposition 'although) (establishing propositional meaning at the sentence level), evidence that she was able to enrich the proposition of sentence 4 and has built a mental model of the first half of paragraph B. In contrast to participant 3, this suggests that establishing a mental model of the paragraph is required to answer item 27 confidently.

Participant 3 selects heading 2 for item 28 (paragraph C) (correct). She references the final two sentences of the paragraph, staying that death is necessary to maintain evolutionary balance ('it talks about how age and death makes possible the biological systems to maintain… balance'; establishing propositional meaning at the sentence level). Participant 6 selects heading 7 (incorrect). Participant 6 completed this item last as he found paragraph C difficult to process. He is unsuccessful and claims that he guesses on the basis that paragraph 2 mentions 'repairs' and that this is also mentioned in heading 10 ('repair of genetic material') (lexical access). No further evidence of greater engagement with this

paragraph is presented. Therefore, establishing propositional meaning at the sentence level is sufficient to address item 28 based on the data from the two participants.

The remaining items were all answered correctly by both participants. For item 29 (paragraph D), participant 3 states that she is aware that paragraph D talks about 'life-span' (lexical access), and she selected option 1 on the basis of the final sentence in paragraph D; organisms' lifespans differ form a few hours/days to several thousand years (establishing propositional meaning at the sentence level), although she quickly realises that this is insufficient data to make the link to heading 1, so deselects option 1, meaning to come back to it later. She returns to answer item 29, selecting heading 7, but in the interview does not offer a detailed explanation of why she did so. This behaviour suggests that processing above the sentence level is required for item 29. Participant 6 also selects heading 7. By way of explanation, he cites and paraphrases sentences 3, 4 and 5), and links them together, demonstrating that he has built a mental model of this paragraph. The heading ('stable life span despite improvements') requires test takers to combine evidence from these three sentences to adequately account for the main and subsidiary noun phrases.

Participant 3 then selects heading 1 for item 30 (paragraph E). She quickly establishes that this paragraph relates to the biological clock on the basis of frequent references to an internal 'clock' and how this 'measures and controls the ageing process', demonstrating an ability to establish propositional knowledge at the sentence level, as the information cited is located in different clauses:

> *"In this paragraph [E], it says 'it is logically necessary to propose the existence of an internal clock', and then it talks about how this clock 'measures and controls the ageing process'. That's why I feel like this paragraph is about 'the biological clock'"* [participant 3, test 1, items 27-32].

Participant 6 also selects the same heading, citing the noun phrase 'internal clock' (lexical access). That he went back to the headings after reading this noun phrase suggests that he was able to hold heading 1 in his working memory. Although he returns to paragraph E to

check his understanding, he does not provide further explanation or change his answer, suggesting that lexical access was not sufficient. Therefore, this item requires establishing propositional meaning at the sentence level:

> *"Because I saw the word 'internal clock', and this paragraph is talking about the 'function' of the internal clock. It's an introduction of the biological clock. I choose the first heading"* [participant 6, test 2, items 27-32].

Participant 3 then considers the final two paragraphs. She moves between the text and the list of headings before selection heading 8 for item 31 (paragraph F). She relates the discussion of 'energy consumption' to 'longevity', demonstrating at minimum an ability to establish propositional knowledge at the sentence level. Participant 6 spends approximately 90 seconds reading this paragraph carefully. He states that this is easier to access on the basis that the topic is familiar and interesting to him and the lexis is familiar. He underlines a noun phrase ('metabolic rate') and summarises the main point of the paragraph ('more active organisms have higher metabolic rate, and [will] die quicker') from multiple sentences, evidence that he has built a mental model of paragraph F. He also successfully summarises the multi-clause sentence 4 (establishing propositional meaning at the sentence level). Sentence level propositional knowledge was sufficient to respond correctly, although participant 6 nonetheless displayed evidence of higher-level processing:

> *"I think the whole paragraph is talking about the 'metabolic rates' of organisms. But, organisms have higher metabolic rate… no, more active organisms have higher metabolic rate, and [will] die quicker. I think the paragraph is talking about 'metabolic rates' and it implies that animals that consume energy more quickly will… the lifespan of them will be shorter, so I think the general idea is about 'energy consumption' [option 8] of animals and organisms"* [participant 6, test 2, items 27-32].

For item 32 (paragraph G), participant 3 selected option 4. By way of explanation, she offers an understanding of paragraphs F and G, demonstrating that she has built a mental model of these two paragraphs:

> *"This paragraph follows up from the previous paragraph, the previous paragraph talks about how energy consumption has an impact on the length of your life, and this one talks about specific strategies for saving energy so that you can achieve a longer life"* [participant 3, test 1, items 27-32].

She combines an understanding of the two paragraphs. Noting that title 4 is personalised ('prolonging your life') she understands that the paragraph will contain personal advice which the reader may apply to their own situation ('specific strategies… so that you can achieve a longer life'). She has gone beyond sentence-level understanding to offer this explanation. Participant 6 summarises the final two sentences of the paragraph [building a mental model]. The limited time spent on this item reflects the shorter paragraph and the personal pronoun in heading 4 linking directly to the personal pronoun used in paragraph G:

> *"Paragraph G suggests that if you make and develop an energy saving programme, you can prolong their life… longer, so I choose 4. It will extend your life"* [participant 6, test 2, items 27-32].

The records of participants 3 and 6 have provided an accurate representation of the requirements for the 'matching heading' item type. This is clearly an item type that targets a range of higher and lower level processes, reflected in the cognitive specification. Participants formed mental models of the paragraphs based on the propositional content. Answering confidently required participants to ensure that all parts of a prospective paragraph heading are addressed within the paragraph and that it encompasses the main point of that paragraph.

4.5.1.7. Type 8: Sentence completion (8-SC)

Strategy code	Strategy label	Freq. rate
11	Careful local reading (text)	2.14
18	Returns to the question for clarification: rereads question and/or options	1.43
19	Searches for key word/phrase (text)	1.29
12	Marks/notes key noun phrase in the text during careful reading	1.00
6	Marks/notes key noun phrase(s) in the questions stem or options	0.86
27	Compares question stem/option to a portion of the text	0.71
2	Identifies the purpose of the question	0.57
1	Reads question(s) before proceeding to the text	0.43
13	Marks/notes key verb phrase in the text during careful reading	0.43

Table 4.29. Most commonly-used strategies for IELTS sentence completion items

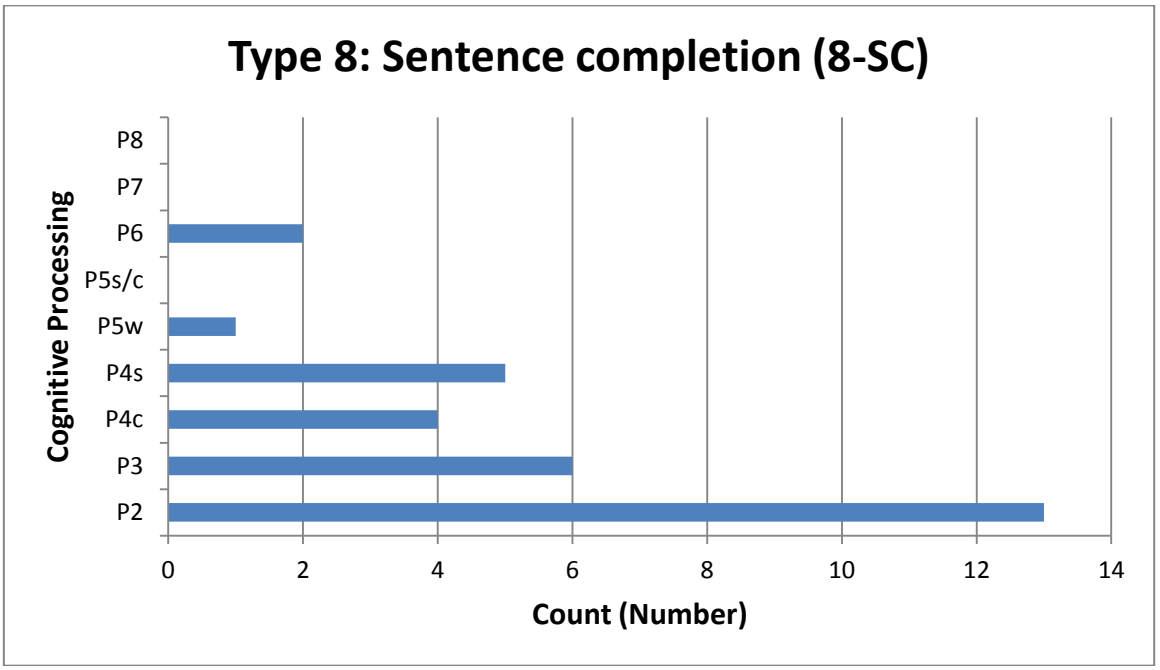


Figure 4.9. Frequency of cognitive processes for IELTS sentence completion items

	P2	P3	P4c	P4s	P5w	P5s/c	P6	P7	P8
Total processes	13	6	4	5	1	0	2	0	0
Processing ratio	1.86	0.86	0.57	0.71	0.14	0.00	0.29	0.00	0.00

Table 4.30. Most frequently-identified cognitive processes for IELTS sentence completion items

Sample items
Questions 4—6

Compete the sentences below.
*Choose **NO MORE THAN TWO WORDS** from the passage for each answer.*
Write your answers in boxes 4—6 on your answer sheet

4 EPRI receives financial support from ……
5 The advantage of the technique being developed by Diels is that it can be used ……..
6 The main difficulty associated with using the laser equipment is related to its …….

Questions 33—36
Complete the notes below.
*Choose **NO MORE THAN TWO WORDS** from the passage for each answer.*
Write your answers in boxes 33—36 on your answer sheet.

- Objects age in accordance with principles of **33**…… and of **34**……
- Through mutations, organisms can **35**…… better to the environment.
- 3**6**…… would pose a serious problem for the theory of evolution

'Sentence completion' items require participants to complete the sentences provided with words taken from the text. Therefore, each sentence is intended to be an accurate paraphrase of one sentence in the text. The sentences will relate to detailed points within paragraphs rather than broader arguments. Test takers will therefore be required to search the text for the relevant parts to answer each item. Items 4-6 relate to text 1, and were completed by participants 1 and 4. Items 33-36 relate to text 3 and were completed by participants 3 and 6. Participants were largely successful with this item type. Participant 1 responded incorrectly to item 6. Otherwise, participants 3, 4 and 6 answered the remaining items correctly.

One observable strategy for this item type was rated 'very high'; 'careful local reading'. Three other strategies were rated as 'high'; returning to the question for clarification (rereads question and/or options); searching for key word/phrase (text) and marking or noting a key noun phrase in the text during careful reading. Item completion procedures then followed a particular pattern. Participants identified key words in the item stem, held the in their working memory and attempted to find identical or similar items in the text. This item type prioritises search strategies, as participants display a significant amount of close engagement with the text in order to find the relevant parts. As sentences will contain paraphrases of parts of the text rather than repeating key words, close reading was required

to confirm that specific parts of the text related to each item stem. Participants tended to mark key words once they identified them so that they could then move between the test and the item seamlessly in order to identify how the part of the text had been restated.

Cognitive processing ratios indicated that this item type targeted lower level processes. The most common process was lexical access, reflecting the importance of using key terms to identify relevant parts of the text. Once identified, engagement was local (at the sentence level) as the items restated specific parts of the text. As close reading of multiple parts of the text occurred, occasional moments of higher level processing were evident in participant verbalisations. Participant 1 began by reading all item stems for items 4-6. She quickly progresses through the three items, identifying the key words (all noun phrases) for the items. She went immediately to paragraph 3 word-matching ('financial support' to 'funded'). This is evidence of the importance of lower-level processing (lexical access). To answer the question, the participant was able to draw on multiple points of reference; matching nouns between the item stem and the text and establish propositional meaning at the sentence level, by understanding that the relative clause in paragraph 3, line 5, refers back to the 'EPRI', the organisation which receives funding from power companies:

> *"It's easy to read, in the previous reading, I know there's some information about finance, so it's easy [to identify the main point]. I know…'EPRI' [Question 4]… and 'financial support', I know there is some support here [paragraph 3], so I directly go to here… financial support, 'which is funded' [paragraph 3]… two words to satisfy the requirements"* [participant 1, test 1, items 4-6].

Chronologically, participant 4 answered items 4 and 5 prior to responding to item 3 (a multiple-choice item). This is due to encountering relevant content in the text relating to these items. Item 3 specifically references the Universities of New Mexico and Florida. The verbalisation demonstrates that the participant initially attempted to answer this item ('key word to question 3'). At 4.38 onwards, the participant identifies the key words 'Florida' and 'New Mexico' in the text (lexical access) before reading around them, but then at 5.53 underlines 'power companies' and writes this in item 4:

> *"Question 4 also has key words. 'EPRI' and the paragraph 3 comes out with the key word, so I think I can do question 4 first. It's obviously the paraphrase of 'support from the EPRI' [paragraph 3]"* [participant 4, test 2, items 4-6].

Correctly responding to this item required an ability to parse a passive structure ('is funded by') and recognise synonymy with the verb phase 'receives financial support from'. The key nouns, 'power companies' and 'EPRI' remain in consistent subject/object positions. Therefore, this question was easily answered using syntactic parsing and establishing propositional meaning at the clause level.

Participant 1 highlighted a name ('Diels') in item 5 as a key noun phrase and used this to locate the specific portion of the text that they needed to access. Once identified, they read ahead of the name to locate the necessary information. The participant displayed evidence that she could parse the sentences in which the correct response is located without difficulty. She then compared the sentence to the stem of item 5 and recognised that she required an adverb for item 5 to be grammatically meaningful (syntactic parsing), and selected the word 'safely' from paragraph 5. This word is located in a different clause from that containing the name 'Diels'; the test taker displayed an ability to establish propositional meaning at the sentence level:

> *"I found somebody's name… I remember that I read it in this paragraph [paragraph 5], so then I go to here… I guess the information may appear in the next sentence, so I used the grammar rules here; 'used [Question 5], maybe I want to find an adverb here. So, while I read this word ['used'], I think it should be an adverb, so I think it is ['safely'] and write it down"* [participant 1, test 1, items 4-6].

Participant 4 proceeded differently. Upon answering item 4, she returned to paragraph 3 to continue reading the information related to the research activities of the two universities in an effort to answer item 3 (multiple choice). Option B suggests that the universities 'use the same techniques' related to lightning research, leading the participant to identify paragraph

5 as important (which includes the name 'Diels'). Whilst examining paragraph 5, she responded to item 5. Paragraph 5 discusses the new laser technique and adds that safety is a major feature of this technique (establishing propositional meaning at the sentence level). Both lexical items, 'safely' and 'safety' appear here. She recognised that she needed the adverb in context (syntactic parsing):

> *"The lines I underlined is the main point of this one from [what] this university [Florida] had done. I can just underline first and then underline it with the next University of New Mexico. I think maybe I should read the next paragraph [5]. I'm getting the point from the one from New Mexico, and then comparing with the [people] from Florida. 'Safely' is in paragraph 5, and it talks about 'safety is a basic requirement', and this one updates this technique by using it more safely"* [participant 4, test 2, items 4-6].

Test takers are required to link the name 'Diels' with the system of discharging lightning he has developed. The characteristic cited in the text is safety. This information is within the same sentence (different clauses. Establishing propositional meaning at the sentence level is required to answer this item successfully.

For item 6, participant 1 is aware that the missing word needs to be a noun (syntactic parsing).

The wording in item 6 requires the test taker to draw parallels between the noun phrase 'main difficulty' and the heading of section 3 of the text 'a stumbling block'. There is no evidence in the text that the participant made this connection, which can be interpreted as insufficient lexical knowledge that led them to read the correct portion of the text. However, their behaviour in relation to the question order and their justification for continuing to access the previous section (*"I know there are still seven tasks I need to answer, but there are only later paragraphs, so I think the answer of this one should be in the previous paragraph"*) suggests that the format of the test was the biggest factor in their decision to read the wrong portion of the text.

Participant 4 uses the subheading on page 2 of the text to recognise synonymy of 'block' and 'difficulty' (lexical access) to identify the relevant portion of the text. She identifies the second sentence of paragraph 7 as significant. The answer is located in the first clause of that sentence and requires the reader to establishing propositional meaning at the sentence level in the question stem. The answer key ('size') is mentioned twice in the second sentence of paragraph 7. The participant circled the second example and cited the first in the verbalisation (*"Paragraph 1 talks about Diels trying to 'cut down the size' and it's some problem with the size, so I think the main difficulty is the size",* participant 4, test 2, items 4-6). Overall, in the first group of 'sentence completion' items (4-6), participants used lower-level processing exclusively to answer the items successfully.

Items 33-36 were completed by participants 3 and 6. Both participants recorded correct answers for all of these items. Participant 3 answered item 36 (prior to answering items 33-35) by writing the word 'immortality'. She is able to verbalise that she remembered this term from paragraph C and is able to elaborate the context in which this word was located. She states that the concept of 'immortality' is problematic in evolutionary terms. She completed this item without reference to the text. This is evidence that she has built a mental model of paragraph C and has established propositional knowledge of multiple sentences and how they interrelate to form meaning at the paragraph level:

> *"I remember I read something about similar about this and I remember there was in paragraph C, when it talks about the issue of ageing and death, and how this is necessary to maintain the balance of the biological system and how immortality would be a threat to the stability of this system, so I immediately thought this should be 'immortality', so I sort of recalled that from my previous reading. 'Immortality' is originally from the text. At this time, I didn't really have to go back to the text to find it out, because I remember I read a similar sentence somewhere in the reading passage"* [participant 3, test 1, items 33-36].

Participant 6 highlights a key noun phrase which he uses to direct him to the relevant portion of the text (lexical access) (*"item 36 mentions the problem of the theory of evolution,*

I still remember there is a sentence that talks about the problem of evolution in paragraph C", participant 6, test 2, items 33-36). He then identifies the subject noun ('immortality') in the penultimate sentence in paragraph C. This is coded as 'syntactic parsing' as the participant admits that he is unsure of the meaning of the term, but recognises that the word fits in grammatically. However, to be certain of this response, he has to link this noun phrase to the pronoun in the subsequent sentence to identify 'immortality' as the 'basic problem of evolution'.

> *"It [item 36] mentions the problem of the theory of evolution, I still remember there is a sentence that talks about the problem of evolution in paragraph C… it's the same. [Paragraph C] said 'immortality' would 'disturb the system', because it 'needs room for new and better life'. And it caused the problem of evolution. So I think 'immortality' is the problem, although I don't understand what does it mean"* [participant 6, test 2, items 33-36].

Participant 6 has demonstrated evidence of inferring at the word level across the relevant sentences, which is the required level of processing to answer confidently. Low level inferential reasoning is therefore the highest level of processing which is necessary for successful completion of a 'sentence completion' item type.

Participant 3 writes 'physical chemistry' for item 33 and immediately after, writes 'thermodynamics' for item 34. For these items, the participant demonstrates that she is able to parse the sentence stem (establishing propositional meaning at the clause level), and links this to paragraph B (lexical access). She recognises synonymy between 'laws' and 'principles', and so copied the subsequent principle noun phrases. The sentence containing items 33 and 34 begins with the word 'objects', which relates directly to the title assigned to paragraph B in item 27, although the participant answered that item incorrectly, suggesting that she did not select this portion of the text on the basis of lexical access alone – she states that she remembered paragraph B contained information pertinent to the principles of ageing:

> *"The sentence here. Here, it talks about the principles of ageing, right? And then I realised that that is mentioned in paragraph B, so that was the paragraph that I went back to. I found the sentence where it actually talks about this, it says 'the laws of physical chemistry and thermodynamics', so because I know this one [items 33 and 34] is about 'principles', so I was sort of reading this paragraph, looking for words like 'principles' and 'laws' and then I found the sentence, so I just copied the words to the text there. I remember the paragraph that talks about [thermodynamics], but I didn't really remember the exact words. I had to go back and find [them]. I read until the sentence where it talks about the laws of ageing, I copied the words and then that's it. I didn't continue"* [participant 3, test 1, items 33-36].

This suggests that she had built a mental model of paragraph B, but that she was required to return to the text as the lexical items required to complete items 33 and 34 relate to supporting details, rather than the main point of the paragraph. She did not continue reading this paragraph once she had answered these items. This high level of processing is not required to successfully complete the item. Participant 6 is able to draw parallels between the wording of item 33 ('principles') and the paragraph content word ('laws') that he held in his working memory (lexical access). Once he had identified the relevant portion of the text, he identified the following noun phrases. He answered item 34 directly after answering item 33. There is no evidence of participant 6 using a level of processing higher than lexical access for items 33 and 34:

> *"I still remember that there is [in paragraph B, mentioned about the 'principles', the 'laws' of physical chemistry and of the more dynamic… I relate them [laws and principles] together [and] objects, the comparison between 'objects' and 'organisms'"* [participant 6, test 2, items 33-36].

Participant 3 returns to the text and focuses her attention on paragraph C. Twenty seconds later, she answers item 35 with the word 'adapt'. She initially uses her understanding of the sentence construction and collocation to identify a plausible word for the gap (establishing propositional meaning at the clause level). She returns to the text to confirm her choice. She

identifies the word ‘adapt’, but cites the following sentence which expands upon the definition of the word ‘adapt’ in the context of evolution. This sentence also contains a paraphrase of the question stem for item 35. To answer this item, the participant was required to combine evidence from two sentences (establishing propositional meaning at the sentence level):

> *“I’m thinking about the next question now. When I was reading this sentence, immediately I want to write the word ‘adapt’, because you know just, sort of without reading, going back to the text I wanted to write down ‘adapt’ because it just feels like this is the word that goes very well with the sentence, but then I went back to the text anyway, just to make sure that… that’s the word, so I went to paragraph C and I got my confirmation after I read this sentence where it talks about ‘tested for optimal or better adaptation to the environmental conditions’”* [participant 3, test 1, items 33-36].

Participant 6 focused on the key words (‘organism’ and ‘environment’) (lexical access) (“I remember paragraph C mentioned the environment, so I just go directly back here”). He identified a sentence that corresponded to the question stem for item 35 (it is almost identical) and therefore chooses the noun ‘adaptation’ and alters the morphology to write in ‘adapt’. There is no evidence offered that he needed to establish the propositional meaning of this sentence to answer correctly – knowledge of sentence structure and morphology was sufficient (syntactic parsing):

> *“I read the question first, and find ‘organisms’ and ‘environment’, I remember paragraph C mentioned the environment, so I just go directly back here. I found a sentence that because of changes in the genetic material, these have new characteristics… they are tested for optimal or better… environmental condition’. [I found the word adapt from] ‘adaptation’”* [participant 6, test 2, items 33-36].

'Sentence completion' item types focus test taker attention on specific details within the text. Occasionally, test takers' knowledge of the text that has developed from engagement with previous items will manifest as a mental model in verbalisations, especially if the answer key for one item is situated in a paragraph which contains a key for a previous item. Mostly, however, this item type requires cognitive processing at the clause and sentence level. As sentences are recasts of parts of the text, test takers can use their syntactic knowledge to identify the missing part of speech which aids them in identifying the relevant part of the text and key.

4.5.1.8. Type 9: Summary completion (9-SuC)

Strategy code	Strategy label	Freq. rate
11	Careful local reading (text)	**1.50**
6	Marks/notes key noun phrase(s) in the questions stem or options	**1.25**
27	Compares question stem/option to a portion of the text	**1.25**
18	Returns to the question for clarification: rereads question and/or options	**1.00**
19	Searches for key word/phrase (text)	**1.00**
12	Marks/notes key noun phrase in the text during careful reading	**1.00**
13	Marks/notes key verb phrase in the text during careful reading	**0.75**
28	Checks/confirms/considers option choice after reading portion of text	**0.75**
1	Reads question(s) before proceeding to the text	**0.50**
21	Marks/notes key phrase in the text during expeditious reading	**0.50**

Table 4.31. Most commonly-used strategies for IELTS summary completion items

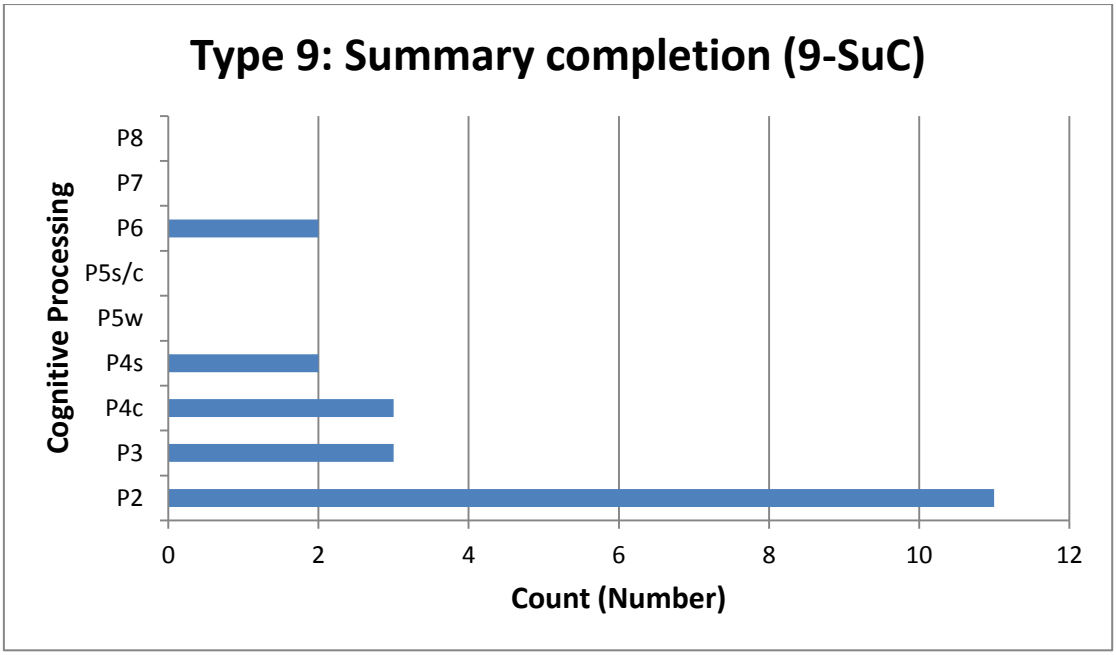


Figure 4.10. Frequency of cognitive processes for IELTS summary completion items

	P2	P3	P4c	P4s	P5w	P5s/c	P6	P7	P8
Total processes	**11**	**3**	**3**	**2**	**0**	**0**	**2**	**0**	**0**
Processing ratio	**2.75**	**0.75**	**0.75**	**0.50**	**0.00**	**0.00**	**0.50**	**0.00**	**0.00**

Table 4.32. Most frequently-identified cognitive processes for IELTS summary completion items

Sample items

Questions 7—10

Complete the summary using the list of words, A-I, below.

Write the correct letter, A—I, in boxes 7—10 on your answer sheet.

In this method, a laser is used to create a line of ionisation by removing electrons from **7**……. This Laser is then directed at **8**…… in order to control electrical charges, a method which is less dangerous than using **9**……. As a protection for the lasers, the beams are aimed firstly at **10**…….

A cloud-zappers
B atoms
C storm clouds
D mirrors
E technique
F ions
G rockets
H conductors
I thunder

The test instrument used in the current study contained four examples of the 'summary completion' item type, all associated with a single summary, items 7-10, linked to text 1. The summary will not relate to the complete text, but may contain both main and supporting details. Therefore, search strategies are required to identify the relevant portion of the text, but once this is located, the location is consistent for all items as the summary may relate to a single paragraph. This item type may necessitate the forming of a mental model, however. As this is a gap-fill type task, test takers will inevitably focus on the surrounding lexis to provide contextual clues to identify the relevant part of the paragraph. Items 7-10 focus specifically on the meaning of a portion of the text. Knowledge of the grammatical structure of the text is less significant as a determiner of success. All of the options are nouns or noun phrases, meaning almost any of the options may go in any of the available spaces. The exception is option E ('technique'), the only non-concrete noun option. All of the spaces in the text are preceded by prepositional phrases (e.g. 'remove from'; 'directed at'; 'aimed at'). The non-concrete noun (technique) also does not appear in the relevant paragraph the test taker must access in order to correctly respond to this section, although it does appear in paragraph 3.

No strategic actions were rated as recording 'very high' usage. However, six strategies recorded 'high' use, indicating a range of approaches undertaken by the two participants. Participants generally use the summary to identify key words and use these to identify the relevant part of the text. They then read this part carefully, returning to the item stems to gain an understanding of how the summary reflects the text. Key words are underlined in both the summary and the text to facilitate this process. This indicates that the majority of processing will occur within sentences.

As the items are associated with text 1, these items were completed by participants 1 and 4. Participants struggled with this item type. Participant 1 answered item seven correctly, but responded incorrectly to items 8-10. Participant 4 responded correctly to items 7 and 9, although answered incorrectly to items 8 and 10. Thus, for this item type, there is a paucity of positive evidence for defining the psychological construct. Negative evidence related to participants' incorrect responses provides evidence regarding an insufficient level of processing, however.

Item 7 is the only example of this item type to be answered correctly by both participants. Participant 1 initially identifies terms to locate the relevant portion of the text. These include noun, verb and adjective phrases (lexical access). She pays specific attention to the phrase 'less dangerous' as it is a comparative, which may assist her in linking the extract to the text (syntactic parsing). She then considers the options before moving to paragraph 6, as she was able to match the lexical items 'removing electrons' from the question stem to 'extract electrons' in paragraph 6 (lexical access). The subsequent noun phrases in the text after 'extract electrons' are 'atoms' (the correct response) and 'ions', both of which she highlights (lexical access). To choose between 'ions' and 'atoms' (options B and F respectively), the participant parsed the clause preceding item 7, recognised the synonymy and related this to 'create ions' in paragraph 6. Thus, a lower-level process (syntactic parsing) was initially used to parse the question stem and the relevant portion of the text, but this was not sufficient to answer the question satisfactorily. Propositional meaning at the sentence level was required for the test taker to be confident of her response. This is also evident in her subsequent elimination of option F:

> *"'Removing' I think is the paraphrase of these words [extract electrons]. I remember at that time, I think I can find these two words, but I have to choose only one, but I think there are two, then I found that it's 'create' not 'removing from' so only one of these is the answer, so at that time, I chose B"* [Participant 1, test 1, items 7-10].

Participant 4 recognises key words from the item stem in the text ('electrons' and 'removing'; examples of lexical access) and recognises the sentence containing the key words as a paraphrase of the first sentence of the summary question (establishing meaning at the clause level). She matches the word 'atoms' from the text with option B:

> *"the question [contains] some key words, 'electrons'… and 'removing', it's just like 'extract electrons'… it's [a] paraphrase, just totally the same… a paraphrase. I mean, 'remove', 'extract' electrons from…"* [participant 4, test 2, items 7-10].

After completing item 7, participant 1 continues reading paragraph 6. Approximately thirty seconds later, she underlines 'directed at a mirror' and then five seconds later selects option D for item 8. Item 8 is answered incorrectly. In the verbalisations and video, there is evidence only of lower-level processing. She identifies words used in the question stem ('directed at') and matches this directly with the same vocabulary in the text (paragraph 6, line 5) and selects the next noun phrase ('mirror') (lexical access). Participant 4 also responds incorrectly. She uses the propositional meaning encoded in the question stem preceding question 8 to identify the relevant portion of paragraph 6 (line 5) (lexical access). Reading more carefully around this point, she is able to verbalise that the laser is not directed straight at the clouds, but is directed at clouds via a mirror, establishing propositional meaning at the sentence level, as this information is encoded in a multi-clause sentence. This item is answered incorrectly, but not because of insufficient textual processing. The question stem is potential for producing a misleading response is due to the belief by the participants that the summary is in chronological order. The laser is aimed at a mirror before being redirected at the clouds.

Both participants mistook items 8 and 10, suggesting they are not independent. To successfully answer item 8, the test taker is required to examine not just the preceding words, but also the subsequent verb phrase ('in order to control electrical charges'). This phrase is not used in the text. To adequately parse this information, the reader needs to be aware that removing electrons form atoms would create a positive electrical charge; information which is not explicitly stated either in the question or the text. A question such as this would undoubtedly favour a test taker with background knowledge in this area. Key words in the question stem such as 'control' and 'electrical' also occur in the text within the same sentence as the answer key (storm clouds), although this information is embedded within a complex, multi-clause sentence with these lexical items separated from the answer key by an adjectival clause. These key words also do not appear in the text in the same context as in the stem, whereas the phrase 'directed at' appears in both stem and text. An item design such as this is potentially misleading for the test taker regarding the processing requirements.

After answering item 8, participant 1 returns to paragraph 6 to address item 9. She spends approximately one minute reading this portion of the text before returning to the question stem to re-familiarise herself with the summary. She states that she is attempting to find reference to a comparative structure but cannot locate one [syntactic parsing]. The participant underlines 'lightning conductors' at 5.47, and then less than five seconds later selects option H for item 9. This does not cohere with the text, indicating that she has selected this option on the basis of lexical access. In the interview, the participant once again highlights the sentence with this reference and reads it aloud, understanding its meaning, and then expresses doubt about her response, although during the test there is no evidence of establishing propositional meaning at the sentence or clause level.

Participant 4 responds correctly to item 9. The participant refreshes her memory of the question stem wording before returning to paragraph 6. She also reads ahead of the item (part of the text preceding item 10). She is attempting to identify key words (lexical access). She correctly identifies the comparative clause structure in the question stem (establishing meaning at the clause level) and realises that the laser technique is being compared to the other technique discussed in paragraphs 3 and 5. Textual clues such as 'method' and 'using' direct the participant to the previous method discussed in the text (using rockets), rather than leaving her to infer this from the textual clue 'less dangerous'. Thus, knowledge of the text (building a mental model) is evident in her verbalisations:

> *"[I'm] reading the next question and read[ing] the later sentence, before this answer [before Q10]. I'm finding the key words in the text… Question 9 is kind of compared with the previous method, and the previous method said that they used the rocket, so I just put the rocket [method] to question 9… compar[ing] the two methods"* [participant 4, test 2, items 7-10].

Item 9 therefore requires test takers to build a mental model by integrating information across paragraphs. The method of using lasers to control lightning is referred to as 'less dangerous than…' with item 9 requiring the test taker to identify the other technique discussed in the text (rockets). However, this second technique is not explicitly stated in

paragraph 6, it is discussed in paragraph 3. Additionally, the words 'less dangerous' do not appear in the text, which the test taker searches for at 3.35. Rockets are not explicitly described as being dangerous. The relevant information is located in paragraph 5, in a quote offered by a university professor ('who wants to fire rockets in a populated area anyway? What goes up must come down'). The test taker is required to infer that the technique is unsatisfactory because of the risks involved to people and property. This item requires ostensibly requires two higher-level processes, of which no evidence emerges in the verbalisations. Yet, it is clear from the participant 1's responses to items 1 and 3 that she is capable of these using these higher-level processes to respond to the test. Potentially, as the participant successfully completed item 7, she may have been under the impression that the required level of processing she identified for item 7 would suffice for items 8-10.

Participant 1 answers item 10 very quickly after item 9 and is answered incorrectly. Evidence suggests the participant focused on 'protection' in the question stem. The verbalisation also suggests that the participant has only selected option A ('cloud zappers') on the basis of lexical access. The speed at which she did this suggests that she did not process any additional portion of the text or attempt to reconcile this with the meaning embedded within the question stem. In the subsequent interview, she admitted this and recognised that the response was likely incorrect:

> *"'Conductors'. Where [did] I find it? Ah, maybe I think I chose this one as protection – I want to find something that they used before, I think, so 'the mirror would be protected by placing lightning conductors close by' [paragraph 6], so this one is wrong, maybe"* [participant 1, test 1, items 7-10].

Use of the key word 'protection' actually directs the test taker to the incorrect portion of the text if this is the only visual clue that they use. The word 'protection' in the text is located in a sentence in which 'mirror' is the head noun. If this was the only clue used and served as the basis of selecting option D ('mirrors'), then the test taker would respond correctly, but would not have activated the level of processing designed into the question. Ultimately, the participant is unable to supply an explanation of why she selected option A, so this is marked as 'guessing'.

Participant 4 also answers this item incorrectly, as this item was partially confused with item 8. She also answers item 10 very quickly after item 9, selecting option C ('storm clouds'). She selects quickly on the basis of lexical access. She identifies the noun 'beam' in both the question and the text and chooses the subsequent noun phrase which is also option C. She readily admits uncertainty after employing this low level of processing. As a result, she returns to the text and underlines a new portion of the text; 'at a mirror' (paragraph 6, line 5). This adverbial phrase describes how the laser beam would be aimed at the clouds. The previous sentence includes the requisite information to understand that the purpose of this technique is to protect the mirror ('to stop the laser being struck'), suggesting that a higher level of processing is required, for which there is evidence in her verbalisation ('some kind of protection for the lasers'). She then selects option D for this item (the correct response). However, there is evidence of test method effect here, as the participant then changes her response on the basis that this option has already been selected (*"I think [option] C is already chosen, and the 'beams', are not mentioned before, until this one, so I think I want to change to it"*, participant 4, test 2, items 7-10). She used her knowledge of the item format (options may only be selected once) rather than her linguistic judgement to change the response back to option C and then reaffirms this choice through low-level lexical access.

Despite participants' lack of success in this item type, sufficient evidence was presented to note that the item type targets both high and low level processes. Participants relied upon key words to identify relevant parts of the text, and lexical features of the summary and text to identify answers. However, in participants' explanations, there were clear instances in which participants needed to gather evidence from more than one sentence in order to answer confidently due to lexical similarity (items 8 and 10) and understanding the progression of the summary (items 9-10).

4.5.2. Cognitive processing in TOEFL iBT item types

4.5.2.1. Vocabulary question (BC-v)

Strategy code	Strategy label	Freq. rate

2	Identifies the purpose of the question	**2.14**
5	Reads question stem(s) and/or option(s) carefully	**1.29**
11	Careful local reading (text)	**1.00**
12	Marks/notes key noun phrase in the text during careful reading	**0.86**
1	Reads question(s) before proceeding to the text	**0.71**
27	Compares question stem/option to a portion of the text	**0.71**
13	Marks/notes key verb phrase in the text during careful reading	**0.57**
25	Eliminates option(s) (contradicts key noun/verb phrase/nonsensical response)	**0.57**
6	Marks/notes key noun phrase(s) in the questions stem or options	**0.43**
20	Skimming part of the text for general understanding (expeditious reading)	**0.43**

Table 4.33. Most commonly-used strategies for TOEFL vocabulary questions

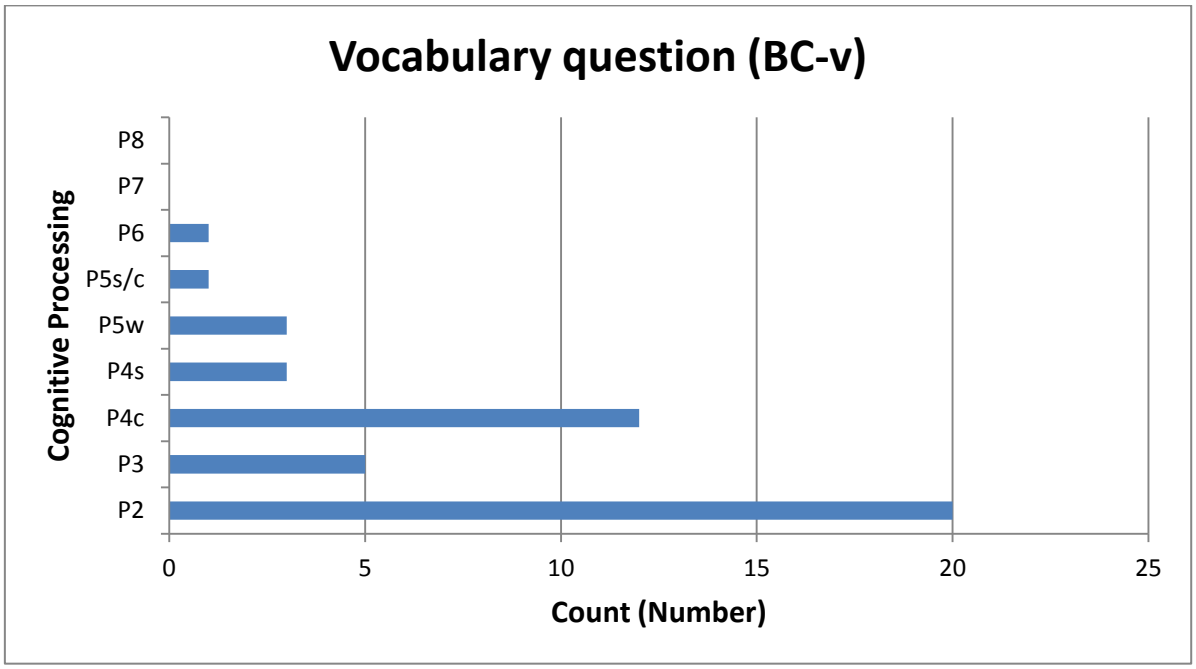


Figure 4.11. Frequency of cognitive processes for TOEFL vocabulary questions

	P2	P3	P4c	P4s	P5w	P5s/c	P6	P7	P8
Total processes	**20**	**5**	**12**	**3**	**3**	**1**	**1**	**0**	**0**
Processing ratio	**2.86**	**0.71**	**1.71**	**0.43**	**0.43**	**0.14**	**0.14**	**0.00**	**0.00**

Table 4.34. Most frequently-identified cognitive processes for TOEFL vocabulary questions

Vocabulary question (BC-v):

> "measure examinees' ability to **comprehend** the **meanings** of **individual words and phrases as used in the context** of the passage" (ETS, 2003: 4, in Cohen and Upton, 2006: 46)

Sample items:

Text reference: "The development of the modern presidency in the United States began with Andrew Jackson who swept to power in 1829 at the head of the Democratic Party and served until 1837. During his administration he **immeasurably** enlarged the power of the presidency."

1. The word **immeasurably** in the passage is closest in meaning to

- frequently
- greatly
- rapidly
- reportedly

Text reference: "Causing participants in experiments to smile, for example, leads them to report more positive feelings and to **rate** cartoons (humorous drawings of people or situations) as being more humorous."

14. The word **rate** in the passage is closest in meaning to

- judge
- reject
- draw
- want

The instrument used for this study contains ten vocabulary items, three related to texts one and three, and four items for text two. Participants 1 and 4 completed items 1, 6 and 8, participants 2 and 5 completed items 14, 16, 22 and 23, and participants 3 and 6 completed items 28, 30 and 33. Evidence from this study shows that participants rely on significantly more than semantic closeness in making decisions about which option to select. Evidence shows that participants use a variety of processes and strategies, and that this is affected by the orthographic form and content of items which require participants to consider options in context.

For this item type, participants generally spend very little time reading the stem and options. Participants rapidly familiarise themselves with this item type and quickly focus attention on the word itself to activate their knowledge from their lexical bank. Participants tended to immediately focus attention on the highlighted word. 'Identifying the purpose of the question' was therefore a highly visible strategy for this item type. As the highlighted words were predominantly nouns, they were highlighted by the participants in both the stem and the text, despite already being highlighted. Options are read through briefly,

focusing on all four for as little as a second. If they are able to understand the highlighted word, selection of the key is almost instantaneous:

> *"[For question 8] I know the 'concept', and 'idea' is the same meaning of 'concept', so I put 'idea' in it [the sentence], it makes sense. The government promoting the general [welfare] is the explanation for the 'concept', so I chose this."* [Participant 4, test 1, item 8]

Participant 4 states that she understands the meaning of the words ('concept') and that it matches the option 4 ('idea'). She states that she mentally places the word in context and then is able to parse it – although she does not state how much of the sentence she parses to determine the word makes sense in context. For this reason, the most common cognitive process coded in the study was word recognition (P2). Item 28 was answered quickly by participant 2, who chose the correct response ('rate') immediately after reading the item and options without further reference to the text:

> *"I guess the four words, for you to read them all takes like 1 second. And 'rate', I know the word, I guess, so I chose 'judge'."* [Participant 2, test 2, item 22]

The participant selected the option without returning to the text to consider any of the words in context. Participant 3 quickly responds correctly to item 30 on the basis of the knowledge that she has gleaned from deep engagement with the text from responding to item 29:

> *"'Relics' means 'remains'. [I didn't consider other options]."* [Participant 3, test 2, item 30]

Only six seconds after answering item 29, the participant selected option 3 for item 30 (correct) without returning to the text. The participant must have used her existing knowledge of the text to respond to this item. She is immediately aware that the highlighted word ('relics') is synonymous with 'remains' (noun) in the context. This may indicate that for

this option, the participant only required 'lexical access' (P2). However, previous verbalisations for item 29 indicate that she had parsed this section very carefully and clearly understood the word 'relics' in context – the previous question related to the age of two mountain ranges and required the participant to infer information about how height was an indication of age. As a result, the participant was able to answer item 30 very quickly without further reference to the text.

However, for instances in which the participant is uncertain of the meaning of the highlighted word and cannot link it directly to a distractor, they return to the text to consider options in context. This was a very common approach for all of the participants. This is a carefully controlled mental strategy which does not appear in the strategy schema as it is not directly observable; the relevant observable actions are 'careful local reading' and 'comparing the question stem to a portion of the text' as the participant moves between text and item. Participants' explanation of this strategy and their resulting reasoning of their option selection accounts for the high proportion of processing occurring at the clause level. A number of examples are evident from participants' verbalisations.

Participant 5 identifies the highlighted word ('rate') in the text (lexical access) and also 'closest', signifying the purpose of the sentence. She moves to the text and underlines part of the text which defines 'cartoons' (lexical access) before returning to consider the options. She moves between the options and the text, comparing them to the meaning of the sentence containing the highlighted word. She circles the verb 'report'; she states that the word 'and' indicates that the words 'rate' and 'report' are fulfilling a similar function (syntactic parsing). She states that 'report' is lexically similar to 'describe' [lexical access] and eliminates 'judge' as she is unaware of the use of the item key in this context, evidence that she has not established propositional meaning. She uses knowledge of collocation and subject knowledge to identify the most likely answer ('draw' cartoons), therefore selects this (incorrect) option (lexical access). This word also ignores the content of the sentence and the function that cartoons have in the research and will differentiate between those participants who have accessed and understood the content versus those who rely on word meaning alone, demonstrating that matching word meaning (lexical access) is not always sufficient to respond correctly to this item type.

For item 30, participant 6 was confounded by the unfamiliar lexis:

> *"I forget what 'relics' means, so I try to find out the meaning of it from the sentence, so I was trying to guess, but I didn't get it, so I just had a guess. Clues [included] 'eroded'. So I think it's related to 'eroded', so I chose the option, which is, I think is related to 'eroded'. I'm not sure about this one! I thought about [the other options], 'regions'… actually I don't know which one is correct."*
> [Participant 6, test 1, item 30]

The participant progresses to item 30, reading the stem and underlining the key word, which he identifies in the text (lexical access). He states that he does not remember the meaning of the highlighted word, so is unable to draw upon his lexical bank for this item. He states that he attempts to identify the meaning of the word in context by linking it to the preceding adjective ('eroded'), although this level of syntactic parsing does not furnish him with the correct response. Ultimately, he guesses option 1 (resemblances) incorrectly.

A number of the distractors contain multiple meanings which the participant will only eliminate from consideration as they progress though the text in a linear fashion. Distractors may be alternative meanings of the word when taken out of context. The highlighted word for item 1 'immeasurably', can be synonymous with option 1, 'frequently', when devoid of context. Conversely, words may have more than one meaning when read independently of context. As participant 1 states in relation to item 1:

> *"I don't know the meaning of this word… 'measurable'. It could be 'calculated', so 'im*measurably' may be [in] opposition [to the] meaning of this word, so I think these three [other options] are not the answers of this question and I put 'greatly' into the original sentence, I think it could be explained logically, so I chose 'greatly'. [I] just put 'greatly' back to*

the sentence [to] check the meaning of the sentence is logical."
[Participant 1, test 1, item 1]

Participant 1 identifies the stem of the word ('measurable') removing both the prefix and suffix and infers (word level) that it can mean 'calculable' and that the prefix gives the word the opposite meaning. This item and response indicates that the test developers intend to allow participants to utilise multiple sources of word knowledge to complete this item type. Participants may access lexical knowledge from their memory bank where knowledge of word meaning is stored literally. Alternatively, meaning may be deduced from orthographic form of the highlighted word. The participant states that the answer she provides ['greatly'] may be explained logically, evidence that she has conceptual understanding rather than offering a linguistic interpretation (e.g. collocation). This suggests that she has established propositional meaning at the clause level. Even if the participants are certain that they understand the connection between the highlighted word and the key from their mental lexical bank, they still return to the text to place the word in context and ensure that their understanding is correct:

"I know, just get the meaning of this word, so there is no need to read the other words. What I do is read the whole sentence, put this word back to the sentence… if I put 'argument' here [in context], it seems very strange. The same with 'example'. I put this word back to the sentence, and I think it's impossible. I can't understand it! Now, I think maybe this one is the best one, maybe just 'tended to be a strong [president]' is to say 'wanted to be a strong [president]'. Yeah." [Participant 1, test 1, item 2]

She eliminates options 1 and 3 on the basis that they do not fit in context, demonstrating that this option cannot be parsed in the sentence (establishing propositional meaning at the clause level). She then provides some metacognitive verbalisations as she reconsiders her choice in the interview, although working through the sentence, she identifies 'inclination with 'tended' and verbally explains that this means 'wanted to be' (establishing propositional meaning at the clause level). As participants mentally place each of the lexical

items in the sentence in place of the highlighted word and reads them in context, it is noteworthy that this strategy is consciously cited, but participants struggle to explain their subsequent selection. Eliminations are made because they "seem impossible":

> *"I put the [options] 'reality' and 'difficulty' into this sentence, and it seems impossible. I can't explain [why]… I think these two are not the answers" [Participant 1, 01:04:07:1-13].*

This is coded as *establishing propositional meaning at the clause level*, as rejecting lexical items must rest upon recognising that a particular word does not cohere in context, even though the participant does not state specifically how choosing each of these items leads to a lack of coherence. As each of the choices in the vocabulary items is the same part of speech (nouns in the case of item 8), relying upon syntactic parsing to recognise that a word does not cohere is not a viable strategy.

4.5.2.2. (Negative) Factual information (BC-nf/BC-f)

Strategy code	Strategy label	Freq. rate
12	Marks/notes key noun phrase in the text during careful reading	**2.92**
6	Marks/notes key noun phrase(s) in the questions stem or options	**2.00**
5	Reads question stem(s) and/or option(s) carefully	**1.83**
11	Careful local reading (text)	**1.67**
2	Identifies the purpose of the question	**1.50**
18	Returns to the question for clarification: rereads question and/or options	**1.25**
1	Reads question(s) before proceeding to the text	**1.00**
27	Compares question stem/option to a portion of the text	**1.00**
13	Marks/notes key verb phrase in the text during careful reading	**0.92**
25	Eliminates option(s) (contradicts key noun/verb phrase/nonsensical response)	**0.92**
28	Checks/confirms/considers option choice after reading portion of text	**0.92**
8	Marks/notes key adjective phrase(s) in the question stem or options	**0.75**
7	Marks/notes key verb phrase(s) in the question stem or options	**0.50**
24	Identifies paraphrase within text or between text and item stem	**0.50**
26	Eliminates option (s) (no information found)	**0.42**

Table 4.35. Most commonly-used strategies for TOEFL (negative) factual information questions

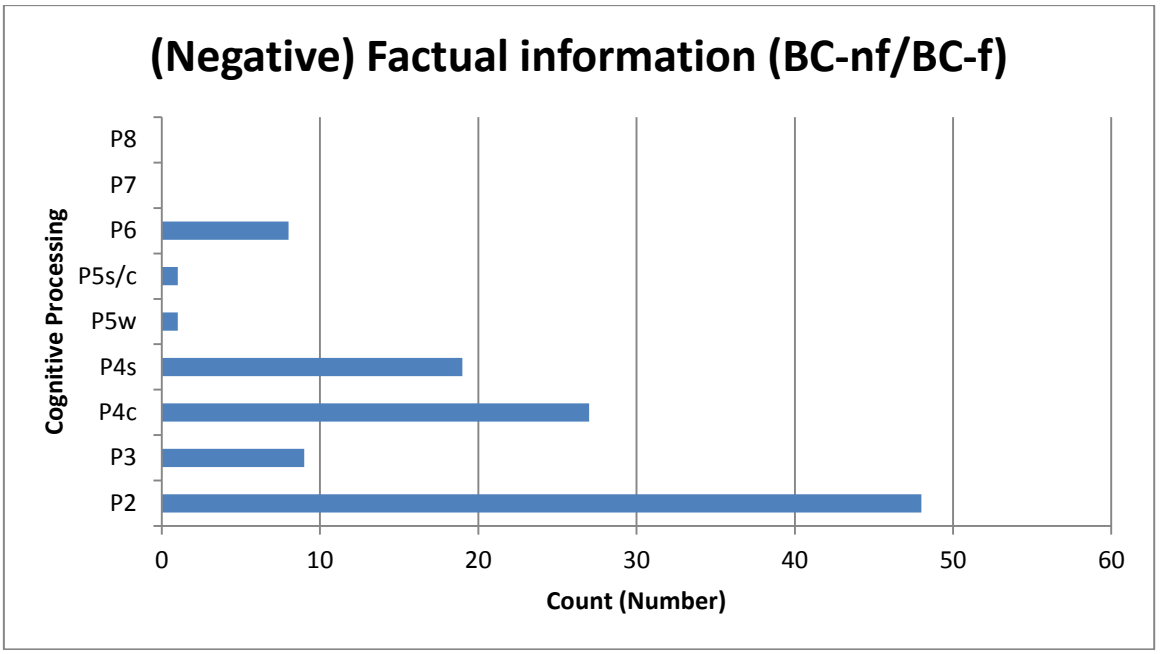


Figure 4.12. Frequency of cognitive processes for TOEFL (negative) factual information questions

	P2	P3	P4c	P4s	P5w	P5s/c	P6	P7	P8
Total processes	**48**	**9**	**27**	**19**	**1**	**1**	**8**	**0**	**0**
Processing ratio	**4.00**	**0.75**	**2.25**	**1.58**	**0.08**	**0.08**	**0.67**	**0.00**	**0.00**

Table 4.36. Most frequently-identified cognitive processes for TOEFL (negative) factual information questions

Negative factual information (BC-nf):

"measure examinees' ability to verify what information is true and what information is NOT true or not included in the passage based on information that is explicitly stated in the passage. The examinees' task is to locate the relevant information in the passage and verify that 3 of the 4 options are true and/or that one of them is false" (ETS, 2003, 12 in Cohen and Upton, 2006: 68-69).

Factual information (BC-f):

"measure examinees' ability to identify responses to questions about factual information that is explicitly stated in a text. The examinees' task is to match the information requested in the item stem to the information in the text that answers the question" (ETS, 2003, 10 in Cohen and Upton, 2006: 63).

Sample items (factual information, BC-f):

20. According to the passage, what did Darwin believe would happen to human emotions that were not expressed?

- They would become less intense.
- They would last longer than usual.
- They would cause problems later.
- They would become more negative.

21. According to the passage, research involving which of the following supported the facial-feedback hypothesis?

- The reactions of people in experiments to cartoons
- The tendency of people in experiments to cooperate
- The release of neurotransmitters by people during experiments
- The long-term effects of repressing emotions

This is the most common item type in the TOEFL test. The test used in this research contains a total of 38 items, of which eleven are factual information questions (and one is a negative factual information question). As there was only a single example of the negative factual item type, it was analysed concurrently with the factual information items, although it is discussed in the next section. The definitions for both item types are offered above. Factual information items require test takers to identify explicitly-stated information in a text and relate this to a paraphrase of this information in the item key. Negative factual information item types differ in that they explicitly require test takers to consider each of the options as an exercise in verification to determine which of the statements contains information that is not accurately represented from the text.

The definition of the factual information item type offered by Cohen and Upton (2006: 68-69) specifically includes matching information between the item stem and the relevant portion of the text. The definition of this item type therefore prioritises local search strategies. This is corroborated by the very high frequency of strategies 6 and 12; marking or noting key noun phrases in the stem and text as participants draw parallels between the two. It was therefore expected that this item type would involve a high proportion of lexical access (P2) processing.

Evidence reveals that participants focus on nouns to identify relevant portions of the text rather than other parts of speech. Marking verb and adjective phrases recorded only medium frequency. The high proportion of word matching strategies stated by all

participants when engaging with this item type translates to a very high frequency rating for 'lexical access' (4.00). 'Establishing propositional meaning at the clause level' also records a very high frequency rating (2.25) as participants then engage with relevant portions of the text.

Verbalisations in which they demonstrate propositional understanding of certain portions are coded accordingly. 'Establishing propositional meaning at the sentence level' records 'high frequency' (1.58). In discussing the factual information items, a selection of responses form participants that highlight the most frequent cognitive processes related to this item type will provide the evidence of the cognitive profile presented in the specification. Verbal records of participant 1 (item 5), participant 2 (item 18), participant 3 (item 27) and participant 4 (item 2) illustrate these cognitive processes and how they were identified. These items are reproduced here for reasons of accessibility.

In responding to item 5, participant 1 displayed evidence of lexical access (P2) and establishing propositional meaning at the sentence level (P4s). She answered this item correctly:

5. According to paragraph 3, which of the following describes the Whig Party's view of the role of government?

- To regulate the continuing conflict between farmers and business people
- To restrict the changes brought about by the market
- To maintain an economy that allowed all capable citizens to benefit
- To reduce the emphasis on economic development

Text reference: "Whigs, on the other hand, were more comfortable with the market. For them, commerce and economic development were agents of civilization. Nor did the Whigs envision any conflict in society between farmers and workers on the one hand and business people and bankers on the other. Economic growth would benefit everyone by raising national income end expanding opportunity. The government's responsibility was to provide a well-regulated economy that guaranteed opportunity for citizens of ability."

> *"Farmers and business people, I think there is no comparison… between the conflict, so it's of course not the answer. I read it again, and I can't find any information about 'restrict the changes', so… and before that, I still can't find something about this one… Option 3 is the most matched information, with*

> *the original passage. The meaning is very similar. The 'all' and 'guaranteed'… yeah. Which words? I think the understanding of this sentence helps me to choose this one. [I matched the words 'capable citizens' to 'citizens of ability']"* [Participant 1, item 5, test 2].

The participant read the final sentence for paragraph 3 and drew brackets around the final clause ("there was a lot of information, so I put [brackets to identify] it more easily"). She then returns to item 5 and eliminates option 1. She states that the content of this option directly contradicts the meaning of the 'conflict' stated in the text. The participant linked information from the stem, the option and the text (establishing propositional meaning at the sentence level). She eliminated option 2 as there is no relevant information for the phrase 'restrict the changes'. This is coded as lexical access, as there is no information available that the participant did anything other than search for the key words in relation to option 2. The participant then selects option 3 on the basis that it contains the "most-matched information", recognising that option 3 and the final (multi-clause) sentence of the textual reference are virtually synonymous (establishing propositional meaning at the sentence level). P2 and P4c are also evident in how participant 2 responds to item 18:

> *"I can tell you why I made this choice… it starts with 'all groups, including the Fore, who had almost no contact with Western culture, agreed on the portrayed emotion', that is obvious I think… synonymous with option 3. I spent like two seconds to read [the other options], I caught the key words, like in the first option I caught the 'shown photographs', the second one, storytelling skills, nothing related to that from that sentence, the third one, 'expression'… 'encourage'… I didn't see 'encourage'"* [Participant 2, test 2, item 18].

Participant 2 spends approximately 45 seconds answering item 18. She cites the sentence *'All groups, including the Fore, who had almost no contact with Western culture, agreed on the portrayed emotions'* as important in her decision. She states that she did not read all of the options carefully before making her selection. She selected option 3 on the basis that it is synonymous with option 3. This was coded as 'establishing propositional meaning at the

clause level' as the information relevant to answering the item is contained entirely within relative clause 'no contact with western culture'. She did not spend significant time relating additional options to the text or engaging with any other part of the sentence. She rapidly identified key words in other options and dismissed them on the basis that the key terms did not appear in the text (lexical access).

The rationale for this item type is testing understanding of propositional information in the text at the clause or sentence level. However, participants 3 and 4 displayed higher-level processing (P6; building a mental model of the text) in relation to basic comprehension items, specifically item 27 (text 3) for participant 3 and item 2 (text 1) for participant 4. Participant 3 initially begins the TOEFL test by engaging closely with the text to get a clear understanding of the themes and argument before moving to the questions. She spends approximately six minutes with the text before moving to page 2. She spends approximately 20 seconds reading the stem and options for item 27 before selecting option 4 (correct). The way that the participant engages with question 27 is especially interesting, as she uses her full knowledge of the text rather than just referring to paragraph 1 to eliminate options 2 and 3:

> *"…all of the other statements* [options 1-3] *were not correct. The first one, it says 'occur more often by uplift than by erosion'. From what I understood about the article, the author didn't really talk about that, and he actually says it is always a continual battle between the two, so they never mentioned whether it was more by this [uplift] or by that [erosion]. 'They occur only at special times'* [option 2]*, the article didn't mention that. 'They occur less frequently now than they once did', again, there was no mention about this either. And then the fourth one, they said that 'they occur quickly in geological terms', there is this indication from the article, it says the Earth is a 'dynamic body', so they do change, maybe slow on human time [term], but granted, compared to the age of the Earth…"* [Participant 3, test 2, item 27]

When selecting option 4, she hesitates for 12 seconds before selecting it. Her explanation suggests that she carefully considered each of the options in turn and related them directly to her mental model of the text that she is holding in her working memory ('from what I understand about the article…'). The participant cites her understanding of paragraph 4 as her reason for eliminating option 1 ('…continual battle between the two…'; evidence of propositional understanding at the sentence level). Options 2 and 3 are eliminated on the basis that no information is contained in the text that relates to them, based on her constructed mental model of the text; she understands that the text does not include information relating to these topics without further textual reference. However, in selecting option 4, she cites the phrase 'dynamic body' in line 1 of paragraph 1, indicating change, and links this to the subsequent sentence that this change is slow for humans, but quickly compared to the age of the earth (establishing propositional meaning at the sentence level). Her explanation indicates that understanding the propositional meaning of the first sentence of paragraph 1 is sufficient to respond successfully to item 27, without citing her knowledge of the text

In contrast, the global knowledge of the text did help participant 4 respond to item 2. Initially she selected option 2 (incorrect) on the basis of lexical access:

> *"I think I'm not sure which one to choose because the first paragraph, I just don't know the whole contents, so I can't get the understanding quickly. [I chose option 2] because this sentence, 'he lectured the senate when it opposed him' and that's how I think it's the point, but later…"* [Participant 4, test 1, item 2]

She examines the stem and options of item 2 and identifies the main purpose of the question, underlining a key adjective phrase. This also demonstrates that she is uncertain of the topic area as she does not underline or highlight key noun phrases (e.g. Andrew Jackson). She selects option 2, but does not provide explicit reasoning for this selection, beyond the synonymy of 'lectured' and 'addressed' (lexical access). Later, she returns to item 2 after having compete the remaining item in the testlet. She states that since she now has a clearer understanding of the text content, she can check her responses:

"Because to look through this article, I think number 2 is especially significant, so I think this one, 'Andrew Jackson' is important more than [the] presidency, so I chose to change my mind. It was the beginning, so because of these people [in the Senate], he started this modern presidency, so I think that's the meaning" [Participant 4, test 1, item 2]

She underlines 'modern presidency' in line 1 [lexical access], which she had not provided evidence of having previously considered – evidence that she has turned her attention in this paragraph to the overall meaning rather than just sentence units. Approximately 50 seconds later, she eliminates option 2 that she had previously selected and instead selects option 3 (correct). She states that her initial understanding of the text centred on the president himself (Andrew Jackson) rather than the office of president – her understanding of the paragraph had clearly developed by this point. She selected option 3 on the realisation that Jackson made a change to the presidency on the basis of his relationship to the senate. Her explanation rules out the possibility that she selected this option due to linking the phrase 'modern presidency' in option 2 and line 1 in the text. Instead she shows evidence of building a mental model of the paragraph by integrating information from multiple sentences.

4.5.2.2.1. Negative factual information (BC-nf)

The test contains a single negative factual information item, reproduced below:

10. According to paragraph 8, the Democrats were supported by all of the following groups **EXCEPT**

- workers unhappy with the new industrial system
- *planters involved in international trade
- rising entrepreneurs
- individuals seeking to open the economy to newcomers

Text reference: "Whigs appealed to planters who needed credit to finance their cotton and rice trade in the world market to farmers who were eager to sell their surpluses, and to workers who wished to improve themselves. Democrats attracted farmers isolated from the market or uncomfortable with it, workers alienated from the emerging industrial system, and rising entrepreneurs who wanted to break monopolies and open the economy to newcomers like themselves."

The negative factual item (item 10) relates to paragraph 6 of text 1 in the TOEFL test. Because each of the distractors has to accurately relate to some aspect of the text, the item covers detailed content rather than the main topic sentence of the paragraph. The paragraph discusses the difference between two political parties in the 19th century United States and the differences in support they receive. Each of the options represents a different group. The test takers must identify which of the cited groups does *not* support the Democrats. This item was completed by participants 1 and 4. Participant 1 responded correctly, while participant 4 responded incorrectly. Both participants begin by highlighting a key word in the item stem ('supported'). They both then turned their attention to the options. Both eliminated option 4 first and for similar reasons; participant 1 notices that 'newcomers' is the indirect object of the sentence in option 4. Participant 4 does not find 'individuals' in option 4 referenced in the text. At this point, both participants seem to be relying on lexical access, as neither link the phrase 'newcomers' to option 3 and note that the 'individuals' who are keen to open the economy to newcomers are actually the 'rising entrepreneurs' cited in option 3 (see text reference above). Participant 1 then notes that in sentence 4 in paragraph 6 (the first sentence cited in the textual reference above) that 'planters' and 'Whigs' co-occur in the same sentence, suggesting that option 2 might be correct:

> *"I read this 'planters', and 'Whigs', I think it might be the answer of this one. This one [planters] is for Whigs, so it's probably not for the Democrats"* [Participant 1, test 2, item 10].

The participant successfully summarises the first clause of the fourth sentence (establishing propositional meaning at the clause level). She links this to option 2 via the noun 'planters' (lexical access), and then explains how option 2 fits the parameters of the task by stating that this group supports the Whigs, so cannot support the Democrats also (inferencing at the clause level). She then eliminates options 1 and 3 as she identifies them as supporters of the Democrats by parsing lines 8-11 in paragraph 6:

> *"It's the supporters of Democrats (option 1). It's the same reason [for option 3] (entrepreneurs)"* [Participant 1, test 2, item 10].

These options are contained within the second sentence in the textual reference above. As these noun phrases are spread across a compound sentence linked by the initial clause 'Democrats attracted…', this was coded as 'establishing propositional meaning at the sentence level'. The participant selects option 2 once she has confidently eliminated options 1, 3 and 4.

In contrast, participant 4 identifies the locations of each option based on the respective head nouns, then returns to the question for clarification by reading the stem and considering each of the options – this is a clearly observable strategy, relating the notion of 'support' in the stem between each of the groups and the Democrat Party. Initially she focuses on the words that she has highlighted (lexical access). She selects option 4 [incorrect] on the basis of lexical access, as she states that the only group cited in the options that are not referred to in the text are 'individuals'. As the question requires her to locate a group that is excluded ('Democrats were supported by all of the following groups EXCEPT…'), she considers this option to be the most likely. She selects an incorrect option as she displays no evidence of processing at a level beyond lexical access. To select the correct option (option 2), she would need to relate the full meaning of option 2 ('planters' and 'international trade') to the fifth sentence (the first in the textual reference included above), which states that planters involved in the world market support the Whig Party. She explains that the penultimate sentence provides evidence for option 4, paraphrasing this option. She is able to link 'entrepreneurs' to 'newcomers' (evidence of establishing propositional meaning at the clause level) but does not link them to the first clause in the sentence ('Democrats attracted…') which would contradict the choice.

Having two participants respond in detail to each item was extremely useful in identifying the level of cognitive processing required to respond correctly to this item. There is clear evidence that the participant who responded incorrectly did so because she did not process at the required level. The participant who responded correctly did so because she processed textual information at the sentence (multi-clause) level, and displayed evidence that she was able to infer information not directly stated in the text. This data verifies the key components of the definition of the negative factual information item type offered by

Cohen and Upton. Participants use the key nouns to locate relevant information. Then verify the content based on "what is not true or not included in the passage". The phrase 'not included in the passage' is clear evidence that inferential reasoning is part of the construct definition for this item type.

4.5.2.3. Pronoun reference question (BC-pr)

Strategy code	Strategy label	Freq. rate
28	Checks/confirms/considers option choice after reading portion of text	**1.50**
30	Hesitates while answering to reconsider choice	**1.50**
12	Marks/notes key noun phrase in the text during careful reading	**1.00**
2	Identifies the purpose of the question	**1.00**
5	Reads question stem(s) and/or option(s) carefully	**0.50**
25	Eliminates option(s) (contradicts key noun/verb phrase/nonsensical response)	**0.50**
20	Skimming part of the text for general understanding (expeditious reading)	**0.50**

Table 4.37. Most commonly-used strategies for TOEFL pronoun reference questions

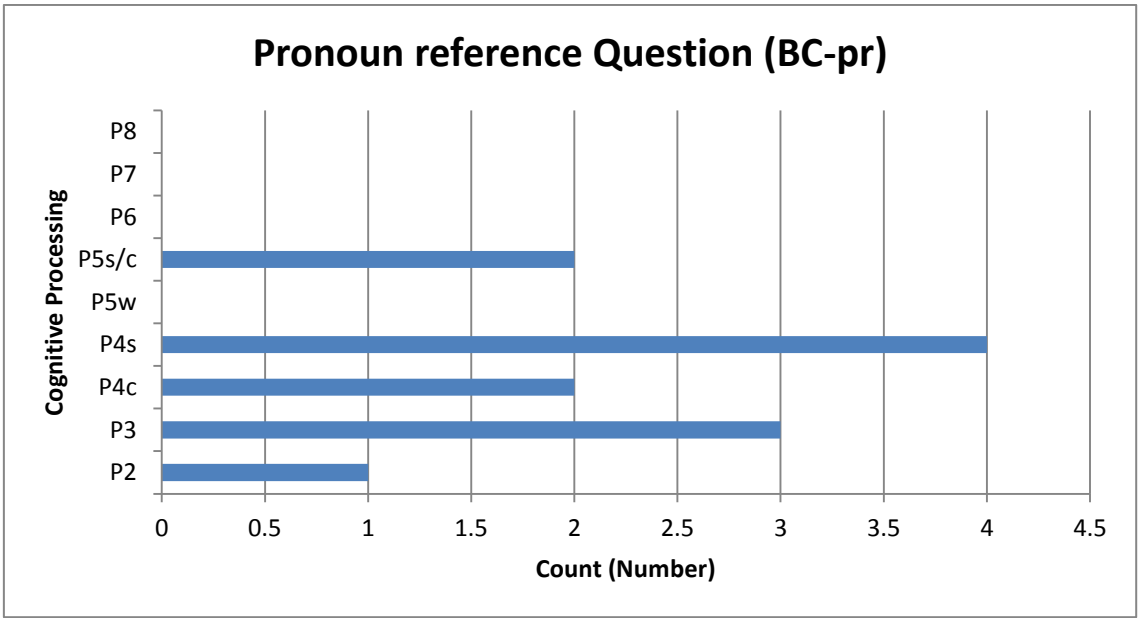


Figure 4.13. Frequency of cognitive processes for TOEFL pronoun reference questions

	P2	P3	P4c	P4s	P5w	P5s/c	P6	P7	P8
Total processes	**1**	**3**	**2**	**4**	**0**	**2**	**0**	**0**	**0**
Processing ratio	**0.50**	**1.50**	**1.00**	**2.00**	**0.00**	**1.00**	**0.00**	**0.00**	**0.00**

Table 4.38. Most frequently-identified cognitive processes for TOEFL pronoun reference questions

Basic comprehension pronoun reference question (BC-pr):

> *"measure examinees' ability to **identify relationships** between **pronouns** and **other anaphoric devices** and their **antecedents/postcedents** within the passage"* (ETS, 2003: 6, in Cohen and Upton, 2006: 51).

Sample items:

Text reference: "Moreover, people in diverse **cultures** recognize the emotions manifested by the facial expressions. In classic research Paul Ekman took **photographs** of **people** exhibiting the **emotions** of anger disgust, fear, happiness, and sadness. He then asked **people** around the world to indicate what **emotions** were being depicted in **them**."

15. The word **them** in the passage refers to

- emotions
- people
- photographs
- cultures

Text reference: "Under very cold conditions, rocks can be shattered by ice and frost. Glaciers may form in permanently **cold areas**, and these slowly moving **masses of ice** cut out **valleys**, carrying with **them** huge quantities of eroded **rock debris**."

34. The word **them** in the passage refers to

- cold areas
- masses of ice
- valleys
- rock debris

Pronoun referencing questions require a test taker to select a single option from four possible options. A consistent item stem is presented to test takers. A pronoun (or other anaphoric reference) is included in the stem and the test taker is required to identify which noun phrase in the text the pronoun refers to. The corresponding pronoun in the text is highlighted to direct the test taker to the relevant portion of the text. As a result, no search strategies were employed by the participants to complete this item. Most common strategies related directly to item completion; participants were careful to consider each of the choices presented to them, skim the text to find them and then mark them in the text. A core strategy is to return to the text, re-read the relevant portion and eliminate options that are not possible. Options can be in the same sentence as the pronoun or up to two sentences away. To avoid word matching, the key should not be the closest noun in more

than 25 per cent of the examples of this item type. In neither item 17 nor item 34 is the key the most proximate noun. As a result, a variety of processing is inferred from participant verbalisations. Across the four participants, twelve specific processes were inferred, the most common being establishing propositional meaning at the sentence level. Evidence of inferential reasoning occurred in relation to item 17 due to the necessity of linking 'photographs' to the verb 'depicted', as shall be seen below.

Two pronoun referencing questions were included in the TOEFL test used in this study. Item 17 was completed independently by participants 2 and 5. Both participants responded correctly. Item 34 was answered by participants 3 and 6, who also both responded correctly to this item. Participant 2 began by reading the options and the question stem. At this moment she explains how her behaviour is guided by the research design, stating that if time were a factor, she would focus on the pronoun ("*Normally, if I know the time is limited, I would just go to the word in question 17"*). However, the research design allowed her time to read the text carefully from the beginning of the second sentence to highlighted word. She spends more than one minute focusing on this portion of the text. She carefully considers option 3 before selecting it. She states that she focused on the sentence preceding the one containing the highlighted word, successfully parses it (syntactic parsing) and identifies the first clause as the relevant unit for analysis. She identifies 'photographs' as the direct object and 'people' as the indirect object (establishing propositional meaning at the clause level):

> *"I see in the sentence, 'in classic research, Paul Ekman took photographs of people', which means this phrase [photographs] is the main word in this phrase. So that is reassuring that I might make the right choice. They are not talking about people; they are talking about [photographs]. Although the people's emotions show in photographs, I think that is the right word."* [Participant 2, test 2, part 2].

Participant 5 spends approximately 25 seconds reading the text, then another seventeen seconds considering option 2 ('people') before selecting it. During this hesitation, the participant states that she is reading carefully. Initially the participant began reading from

the beginning of the sentence containing the pronoun ('He then asked…'). This meant that she did not identify the relevant noun, so initially selected the one that she most closely associated with the highlighted word. She ruled out option 1, on the basis that it refers to clearly-stated example emotions in the previous sentence (syntactic parsing). She eliminates option 4 on the basis that the word is too broad to fit into the context (lexical access). She then associates the word 'photographs' from the previous sentence with the word 'depicted' (inferencing at the sentence level) and identified the deficiency with option 2; it does not fit into the passive sentence structure (establishing propositional meaning at the clause level):

> *"It can't be 'emotions' because the 'emotions' already appeared here, so I already excluded this one, and 'cultures' is a little bit blurred, it's not very close to the meaning, so I think it should be 'people'. I'm not that clear, I didn't really think very carefully on this actually. Because it's 'depicted', it's more like in 'photographs', you can't say depicted by 'people'."* [Participant 5, test 1, part 1].

Item 34 was completed by participants 3 and 6. Participant 3 begins by briefly reading the question stem and options, but does not need to move to the text (paragraph 6) before selecting the correct option (two). In item 34, the highlighted pronoun occurs in a participle clause, indicating that it is not the main clause of the sentence and the reader must link back to the previous clause in order to parse it. The participant correctly links the pronoun with the previous demonstrative pronoun ('these'), which indicates the noun phrase 'masses of ice' is the key noun phrase in this multi-clause sentence (establishing propositional meaning at the sentence level):

> *"'Them' is a kind of continuation from the previous part of the sentence; these slowly-moving masses of ice, carrying with them huge quantities of blablabla. So that means 'them' must be the noun phrase here mentioned before."* [Participant 3, test 2, part 2].

Once she has selected the option, she then moves to the text to review her choice and perform an elimination exercise; option 1 is rejected as it represents a part of speech (syntactic knowledge). The participant states that she compared option 3 ('valleys') with option 2 ('masses of ice') as option 2 was the closest noun to the pronoun, so needed to be considered. She selected option 2 on the basis that the noun ('them') connects with the adjective phrase 'slowly moving' – it is the ice that moves and not the valley. This is evidence of establishing propositional meaning at the sentence level, as the participant unifies and parses information across clauses:

> *"So, 'cold areas'; that's like an adverbial clause, and that means it's not the main part of the information. And 'valleys' is the noun in front of, before this sub-sentence with 'them' in it. I went for the first one because what was moving was the masses of ice, not the valleys. The ice can only cut out the valleys, but could not move the valleys with them."* [Participant 3, test 2, part 2].

Participant 6 begins by reading the text – he reads the whole paragraph before turning his attention to question 34. Only six seconds after he began reading the stem, he selects option 2 (correct). He shares his understanding of the first two sentences, stating that glaciers make valleys. The pronoun links back to the first part of the sentence. He puts 'masses of ice' back into the sentence and it makes coherent sense to him (establishing propositional meaning at the sentence level). 'Them' refers back to the previous clause – masses of ice being the subject of the sentence and synonymous with 'glaciers':

> *"It talks about the glaciers carrying the ice, the 'masses of ice'. And to make valleys, and then, 'them' here is connected to the first part of the sentence, so it mentions that glaciers carry the masses of ice and then I think 'them' is continuing to talk about the masses of ice. I don't know how glaciers can carry valleys, 'rock debris' is… because it's carrying with rock debris, it's not very correct, so I just chose 'masses of ice'. I just put 'masses of ice' into the sentence, replacing [them] to read it again."* [Participant 6, test 1, part 2]

Item 17 mandates the establishment of propositional meaning in more than one sentence and then linking this information together. In this sense, item 17 can be defined as an inferencing item, as not only is the target word is in the previous sentence, but it is also a pronoun referring back to the previous sentence. Khalifa and Weir (2009) provide one definition of 'inferencing' as linking information across sentences at the word level (e.g. pronouns).

One aspect of item difficulty related to pronoun referencing is the distance between the pronoun and the target word. In item 17, there are 27 words between the target word and the highlighted pronoun. In item 34, there are only five. The influence of this disparity is evident in the actions of participant 6. He responds correctly to the item only six seconds after reading the options, without referring back to the text. Conversely, participant 2 spends more than a minute attempting to parse the three sentences cited as the textual reference for item 17. The greater the distance between the pronoun and the referent, the greater the amount of information the item requires the test taker to parse and cohere.

In order to respond successfully to this item type, the participants have to form textual coherence between pronouns or other anaphoric devices and their given referents. In the case of item 17, the anaphor is 'them'; a third person plural object pronoun. The sentence containing the target pronoun also contains the pronoun 'he'. The target pronoun is relatively abstract by comparison. Identifying a gender-specific subject pronoun such as 'he' immediately prompts the reader to infer that the sentence will reveal an action undertaken by a named individual that the reader committed to short-term memory in the parsing of the previous sentence. In the case of item 17, this is 'Paul Ekman'. The reader is further assisted in forming a semantic bridging inference between the two sentences as two key noun phrases in the target sentence contain co-referent noun phrases in the previous sentence; 'people' and 'emotions', which are also two distractors used in item 17. The only remaining noun phrase form the preceding sentence unaccounted for in the target sentence is 'photographs'. The remaining distractor occurs in the sentence before this one.

Both of these items reference processes. Item 17 discusses research design and item 34 addresses a physical process. Item 34 cites a cause-effect process as part of an expository

text. Participants are required to infer the causes of physical outcomes, which nouns are the agents and which are the objects. Three of the four options are located before the target pronoun, with one occurring after. In item 34, the referent for the pronoun is in the same sentence as the pronoun, although is in a different clause. The target pronoun is located in a subordinate present participle clause, so is providing additional information to that carried in the main clause. The key, 'masses of ice' was selected as an option in place of 'glaciers'. The participants must cohere 'masses of ice' with 'glaciers' and recognise that they are synonymous in this context (in a different context, 'masses of ice' could mean 'iceberg'). Participants are also required to understand that 'carrying' implies movement, and so must decide what is carrying what. In the cited text, there is an implicit causal relation between the rocks being broken by ice and frost, and glaciers carving out valleys and carrying rocks with them.

4.5.2.4. Sentence simplification question (BC-ss)

Strategy code	Strategy label	Freq. rate
28	Checks/confirms/considers option choice after reading portion of text	**2.67**
12	Marks/notes key noun phrase in the text during careful reading	**2.00**
2	Identifies the purpose of the question	**1.67**
5	Reads question stem(s) and/or option(s) carefully	**1.67**
1	Reads question(s) before proceeding to the text	**1.67**
25	Eliminates option(s) (contradicts key noun/verb phrase/nonsensical response)	**1.33**
27	Compares question stem/option to a portion of the text	**1.33**
8	Marks/notes key adjective phrase(s) in the question stem or options	**1.33**
26	Eliminates option (s) (no information found)	**1.33**
24	Identifies paraphrase within text or between text and item stem	**1.00**
11	Careful local reading (text)	**0.67**
18	Returns to the question for clarification: rereads question and/or options	**0.67**
7	Marks/notes key verb phrase(s) in the question stem or options	**0.33**
14	Marks/notes key adjective phrase in the text during careful reading	**0.33**
10	Marks/notes key adverbial phrase(s) in the question stem or options	**0.33**
34	Checks progress	**0.33**

Table 4.39. Most commonly-used strategies for TOEFL sentence simplification questions

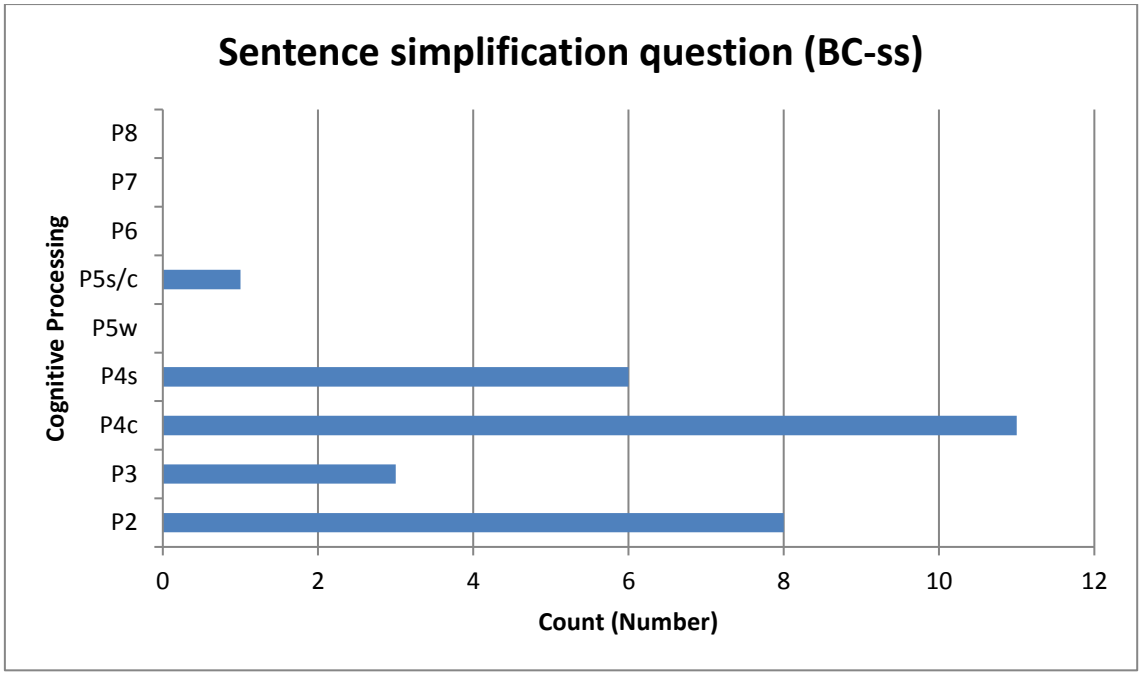


Figure 4.14. Frequency of cognitive processes for TOEFL sentence simplification questions

	P2	P3	P4c	P4s	P5w	P5s/c	P6	P7	P8
Total processes	**8**	**3**	**11**	**6**	**0**	**1**	**0**	**0**	**0**
Processing ratio	**2.67**	**1.00**	**3.67**	**2.00**	**0.00**	**0.33**	**0.00**	**0.00**	**0.00**

Table 4.40. Most frequently-identified cognitive processes for TOEFL sentence simplification questions

Sentence simplification question (BC-ss):

> *"measure examinees' ability to identify essential information as they process complex sentences in extended texts without getting lost in less important details and elaborations"* (ETS, 2003: 8, in Cohen and Upton, 2006: 56).

Sample items:

Text reference: "The Whigs were strongest in the towns, cities, and those rural areas that were fully integrated into the market economy, whereas Democrats dominated areas of semi-subsistence farming that were more isolated and languishing economically."

11. Which of the sentences below best expresses the essential information in the highlighted sentence in the passage?

Incorrect choices change the meaning in important ways or leave out essential information.

- Whigs were able to attract support only in the wealthiest parts of the economy because Democrats dominated in other areas.
- Whig and Democratic areas of influence were naturally split between urban and rural areas, respectively.
- The semi-subsistence farming areas dominated by Democrats became increasingly isolated by the Whigs control of the market economy.
- *The Democrats' power was greatest in poorer areas while the Whigs were strongest in those areas where the market was already fully operating

Text reference: "The Fore also displayed familiar facial expressions when asked how they would respond if they were the characters in stories that called for basic emotional responses."

19. Which of the sentences below best expresses the essential information in the highlighted sentence in the passage?

Incorrect choices change the meaning in important ways or leave out essential information.

- The Fore's facial expressions indicated their unwillingness to pretend to be story characters.
- The Fore were asked to display familiar facial expressions when they told their stories.
- *The Fore exhibited the same relationship official expressions and basic emotions that is seen in Western culture when they acted out stories.
- The Fore were familiar with the facial expressions and basic emotions of characters in stories.

Text reference: "Hills and mountains are often regarded as the epitome of permanence, successfully resisting the destructive forces of nature, but in fact they tend to be relatively short-lived in geological terms."

35. Which of the sentences below best expresses the essential information in the highlighted sentence in the passage?

Incorrect choices change the meaning in important ways or leave out essential information.

- When they are relatively young, hills and mountains successfully resist the destructive forces of nature.
- *Although they seem permanent, hills and mountains exist for a relatively short period of geological time.
- Hills and mountains successfully resist the destructive forces of nature, but only for a short time.
- Hills and mountains resist the destructive forces of nature better than other types of landforms.

Test takers are presented with an item stem directing them to a highlighted sentence in a paragraph. They are presented with four shortened versions of this sentence and must select the one which contains the same meaning as the highlighted sentence. Test takers

are informed that incorrect options either change the meaning or leave out important information. Test takers will need to consider the distractors carefully and identify what information has been omitted to answer examples of this item type confidently.

The most commonly used strategies were 'checking or considering an option choice after reading portion of text' and 'marking or noting key noun phrases in the text during careful reading'. These recorded 'very high' frequency use across the six participants, reflecting the high processing load associated with this item type. Participants are required to determine which of the options best reflects the information in a long, complex sentence. Revisiting the text and considering each option, identifying noun phrases to draw parallels between the options and the sentence are necessary item management strategies. Other frequently-used strategies include 'reading question stem(s) and/or option(s) carefully'; 'eliminating option(s)'; 'comparing question stem/option to a portion of the text'; 'identifying a paraphrase' and 'marking or noting key adjective phrase(s)'. Key adjectives occur in item 11 and item 35, which account for the high frequency of highlighting adjectives as a strategy. These strategies together provide a clear picture of how participants complete this item type. Participants examine the question stem before reading the highlighted sentence carefully. They frequently move between the text and the item, identifying parallel words or phrases (such as nouns and adjectives) and marking them. They eliminate options if they find evidence that it contains information contrary to the highlighted sentence before selecting an option.

Participants' verbal explanations of these strategic actions revealed that the most common cognitive processes were 'lexical access' and 'establishing propositional meaning at the clause and sentence levels'. Despite the apparent complexity of the item type, there was only a single instance of higher-level processing observed. High incidence of lexical access corresponds to strategic actions relating noun and adjective phrases in options to the highlighted sentence. Participants demonstrated syntactic knowledge when they specifically sought out noun phrases in order to read around them. Participants explained why they eliminated or selected options by establishing propositional meaning at the (local) clause level and at the sentence level. One instance of 'inferencing at the clause/sentence level'

was recorded when a participant used her own schematic knowledge to eliminate an option that was not feasible.

For sentence simplification items, each of the highlighted sentences must be grammatically complete. That is, sentences should contain complete ideas that do not require reference to other parts of the text to disambiguate meaning. Distractors should be composed of at least one clause that includes information that directly contradicts affirmative information in the given sentence or provide information that is not included in the highlighted sentence. Options may provide both content and grammatical context clues to give the correct response. For example, in item 35 above, option 3 includes a main clause that coheres with the main clause of the highlighted sentence ('hills and mountains are often regarded as the epitome of permanence'). However, the third clause in the sentence then contradicts this claim. Option 4 provides a comparative statement declaring hills and mountains to be more resistant to the forces of nature than other landforms. Other landforms are not cited in the highlighted sentence, and there is no grammatical comparative structure. The key for this item is option 2:

- *Although they seem permanent, hills and mountains exist for a relatively short period of geological time.*

The noun phrase 'hills and mountains' is repeated in each of the options, and each of the three distractors cites 'the destructive forces of nature', eliminating word matching as an item completion strategy. Crucial information in the highlighted sentence related to the passive verb phrase ('are often regarded as') is contained within the subordinate clause ('although…'). The contrast between the clauses is maintained by the addition of the conjunction 'although' in place of the conjunction 'but'. Evidence emerged of participant 3 using syntactic structure to successfully respond to item 35:

> *"The idea of permanence is used in a noun phrase [in paragraph 2, line 1], but this one is an adjective [option 2]. And then the first sentence, it's, you know they use 'but' [line 2], and in this sentence it used 'although [option 2] but they generally mean the same thing*

> *[syntactic parsing]. This statement is a main argument that the author was trying to develop in the article and it's the same with the… it's simply saying the same thing as this sentence here, so… I didn't really go for the words, I think I just went for the ideas, and I feel like the idea's the same. I think, you know the 'hills and mountains' seem to be permanent, but actually, you know, they exist for a short time, in geological time… number 3 says 'hills and mountains successfully resist', but apparently that's not the case, because no hills and mountains can resist".* [Participant 3, test 2, item 35]

The participant states that in selecting option 2, she based her decision on the ideas in the text rather than the words (establishing propositional meaning at the sentence level), although she is able to explain her choice in terms of textual references. She successfully eliminates option 3 as there is no explicit textual information. Here, the participant shows an ability to infer at the sentence level - as she is bringing her own knowledge of the real world to bear on option 3. She is aware that there is no resistance to the forces of erosion in the natural environment, even for a short time. Participant 6 offers a similar explanation that is equally informed by syntactic and semantic meaning of the highlighted sentence and option 2:

> *"I read the highlighted sentence, to get the idea of it. I think the option 2 is most suitable, is most similar with the meaning of the sentence, so I choose option 2. The sentence says 'hills and mountains are often regarded as permanent, but…'. It's [this] 'but', in fact they tend to be relatively short-lived', so this, option 2 is in the same format, 'although', they are permanent, 'but' 'they exist for a relatively short geological time'".* [Participant 6, test 1, item 35]

He highlights the conjunction 'but' as the discourse marker that indicates 'hills and mountains' are not permanent (syntactic parsing). He notes that option two has the same linguistic organisation; the conjunction 'although' introduces a similar proposition

(establishing propositional meaning at the clause level). The participant then uses understanding of the ideas contained within the text and options to eliminate distractors:

> *"I [read the other options], but the meanings of them are not very precise compared to option 2. When they are relatively young [option 1], hills and mountains successfully resist the destructive forces of nature', it's not the meaning of the highlighted sentence. The highlighted sentence is trying to emphasise the short-lived… but the option 1 does not emphasise that, same as 4. Option 4 is talking about the destructive forces of nature better than other types, but it doesn't mention it in the sentence. And this option 3, 'hills and mountains successfully resist the destructive forces of nature', it means hills and mountains can only resist for a short time, but actually, the highlighted sentence is talking about the hills and mountains seem[s] to be very permanent, but higher mountains tend to be relatively younger, the meaning of option 3 is not the same as the meaning of the highlighted sentence."* [Participant 6, test 1, item 35]

The participant correctly states that option 1 is contradictory – hills and mountains do not resist the forces of nature (establishing propositional meaning at the clause level). The passage highlights that hills and mountains have a short life-span (establishing propositional meaning at the clause level). A similar level of processing is exhibited for options 3 and 4 in recognising that these options are contradictory to the meaning of the highlighted passage [establishing propositional meaning at the clause level]. Option 3 is contradictory as the participant notes that hills and mountains do not resist the forces of nature, even for a short time (establishing propositional meaning at the clause level).

The key is written so that each example of this item type cannot be successfully completed without knowledge of the meaning of the constituent lexical units. Word matching or matching linguistic structures is not a viable strategy for this item type, as is evident from the verbalisations of participant 1 for item 11:

> *"I hadn't found information referred to in the sentence, I can't find the 'wealthiest'. Because in the Whigs, there is 'urban' and 'rural', so it should not be a difference between these two parties. This information… ['semi-subsistence'] and, this is the… 'market economy'. Yeah. Then I just wanted to double-check [option 4]. There is no information about 'poorer' [areas] or [whether the market was] 'fully operating', so I think it's not correct. To be exact[ly], I found this type of question a little bit difficult."* [Participant 1, test 2, item 11]

The participant selected option 3 (option 4 is correct). This is the only item in her TOEFL test that she answered incorrectly. She considers the highlighted sentence and identifies what she suspects will be relevant noun phrases (syntactic parsing). She moves to the question options and follows this strategy of attempting to identify which are most closely matched to the highlighted sentence on the basis of content (lexical access). She focuses on the word 'wealthiest' in option 1 and attempts to identify relevant information in the highlighted sentence. Although different groups are identified as supporting the two political parties, the text does not state whether these groups divide into 'wealthy' and 'less wealthy' sectors of the economy. She also considers option 2, and eliminates this option after identifying the key words 'rural' and 'urban' – she states that the Whigs have both rural and urban supporters. This information is located in the first clause of the highlighted sentence and so this statement is evidence of establishing propositional meaning at the *clause* level. She then selects option 3 after underlining 'semi-subsistence' and 'market economy' in the option stem. She cites her selection of this option on the strength of these two terms (lexical access) without further reference to sentence content. Further evidence that she has not established the propositional meaning of the highlighted sentence is given when she explains why she eliminated option 4:

> *"There is no information about 'poorer' [areas] or [whether the market was] 'fully operating', so I think it's not correct"* [Participant 1, test 2, item 11].

She does not associate the terms 'poorer' and 'fully operating' in option 4 with 'languishing economically' and fully integrated' in the highlighted sentence respectively, which is instrumental in selecting option 4, as demonstrated by deeper engagement with the highlighted sentence by participant 4:

> *"It talks about the comparison of the two groups. The Whigs and the Democrats. It's the comparison of the two groups' influence. I think it [option 1] is not correct, because the sentence is [contains] 'only'. As soon as I looked at the 'only', I thought it was wrong, because the sentence said the word 'strongest', it didn't say that 'only', support only [from the wealthiest parts of the economy]. The context just makes the comparison of two groups, but it didn't say that [farming areas] were isolated by the Whigs. Number 2, I'm not sure. It seems that the sentence paraphrases the bottom three lines of the paragraph. 'Integrated' and 'strongest' and the 'market already fully operating' and the same [is] true [of] 'fully integrated in the market economy'."* [Participant 4, test 1, item 11]

When she considers option 1, she notices that the first clause contradicts the opening clause of the highlighted sentence, and that the words 'only and 'strongest' are not synonymous and that this changes the meaning in an important way (establishing propositional meaning at the clause level). She therefore excludes this option. She moves on to consider each of the remaining options and eliminates option 3 from consideration as it contradicts the content of the highlighted passage; option 3 states that 'farming areas were isolated by the Whigs', which is not included in the sentence. Words are repeated in both contexts, suggesting the participant was able to establish propositional meaning at the clause level. She then considers the two remaining options (2 and 4). Option 2 focuses on the urban/rural split cited in the sentence, although this information is included as an example of areas that were fully-integrated into the market economy versus those that were not. The participant recognises some parallel content words (lexical access). She ultimately selects option 4 and offers examples of parallel lexical phrases rather than the meaning explicitly. Establishing propositional meaning at the clause level by comparing constituent parts of the sentence to the options is sufficient for her to record the correct response.

Evidence that a participant has understood the meaning of the sentence from engaging with the paragraph for previous items came in the form of the response to item 19 by participant 2. She is able to select the correct response to this item within thirty seconds of accessing the question stem:

> *"I'm reading the words in bold letters… I got the general idea of what the sentence is about, it's kind of the relationship between facial expressions and emotional responses, so after that, I came to the four options and after I read them I found [that] the third one is most relevant. The relationship between 'facial expressions' and 'basic emotions' is the key idea of that sentence in bold letters. For the other three options, I don't think they are related; they are biased. I think the third one is more comprehensive, or more synonymous with that one [highlighted sentence]. The only thing that made me a little bit hesitant is 'acted out stories' [option 3], because this is kind of a misleading… I'm not sure about the meaning of 'acted out', but option 3 has got the main idea of that sentence"* [Participant 2, test 2, item 19]

She states that after reading the sentence in bold, she has a "general idea" of what it is about and is able to summarise the sentence in her own words (establishing propositional meaning at the sentence level). She then read the options without referring back to the text and identified the third option as "most relevant" (establishing propositional meaning at the sentence level). However, she does state hesitation in selecting option 3 due to the final three words of the option ('acted out stories') as she is uncertain of the meaning – she is unable to relate it to the hypothetical notion of pretending to be characters contained within the highlighted sentence. But she is able to link 'familiar facial expressions' to those 'seen in Western culture'. In her verbal explanation of completing the item, she refers to ideas contained within the text rather than the words in the options and linking them to the highlighted sentence. Participant 5 offered a very similar explanation:

> *"I choose the third one, because it says the most closely meaning… because these two points, before, we already talked about the Western culture that they are really not close to it, but they still agreed on the portrayed emotions,*

so I think in this second part of this paragraph 2 is just trying to say, no matter about the cultural background, people share the same emotions, expressions. I was [then] considering the other options. The first choice, their unwillingness, I think it doesn't mention in the first paragraph, so I exclude the first one, and then the Fore weren't being asked to do the similar facial expressions, they were just asked what [how] they would respond, so it's not the wrong way to understand that. 'Were familiar', I think in the paragraph, they only said they will copy, maybe, when they are telling stories, when they think they are characters in the story, they… it seems this could be the right answer."
[Participant 5, test 1, item 19]

She states that option 3 is closest in meaning based on the reference to Western culture (lexical access), that the Fore people agreed on the portrayed emotions in the photographs (establishing propositional meaning at the clause level). She then explains that people share the same emotions, regardless of background (summarising the main point of the paragraph given in the first sentence, thereby establishing propositional meaning at the sentence level). She states that she considered the other options – option 1 is eliminated as the paragraph does not reference the participants' willingness to act out stories (lexical access). The second response is eliminated as it contradicts the parameters of the experiment stated in the highlighted sentence (establishing propositional meaning at the clause level). The participant states the Fore were familiar with emotions in the stories, which is true, given they portrayed them, but does not contain information about the Fore being asked to act out stories (establishing propositional meaning at the sentence level).

4.5.2.5. Inference question (I)

Strategy code	Strategy label	Freq. rate
12	Marks/notes key noun phrase in the text during careful reading	**3.00**
5	Reads question stem(s) and/or option(s) carefully	**3.00**
6	Marks/notes key noun phrase(s) in the questions stem or options	**3.00**
11	Careful local reading (text)	**2.50**
28	Checks/confirms/considers option choice after reading portion of text	**2.00**
2	Identifies the purpose of the question	**2.00**

25	Eliminates option(s) (contradicts key noun/verb phrase/nonsensical response)	**2.00**
27	Compares question stem/option to a portion of the text	**1.00**
8	Marks/notes key adjective phrase(s) in the question stem or options	**1.00**
13	Marks/notes key verb phrase in the text during careful reading	**1.00**
1	Reads question(s) before proceeding to the text	**0.50**
7	Marks/notes key verb phrase(s) in the question stem or options	**0.50**
20	Skimming part of the text for general understanding (expeditious reading)	**0.50**
32	Uses own topic knowledge to enhance understanding of text/questions	**0.50**

Table 4.41. Most commonly-used strategies for TOEFL inferencing questions

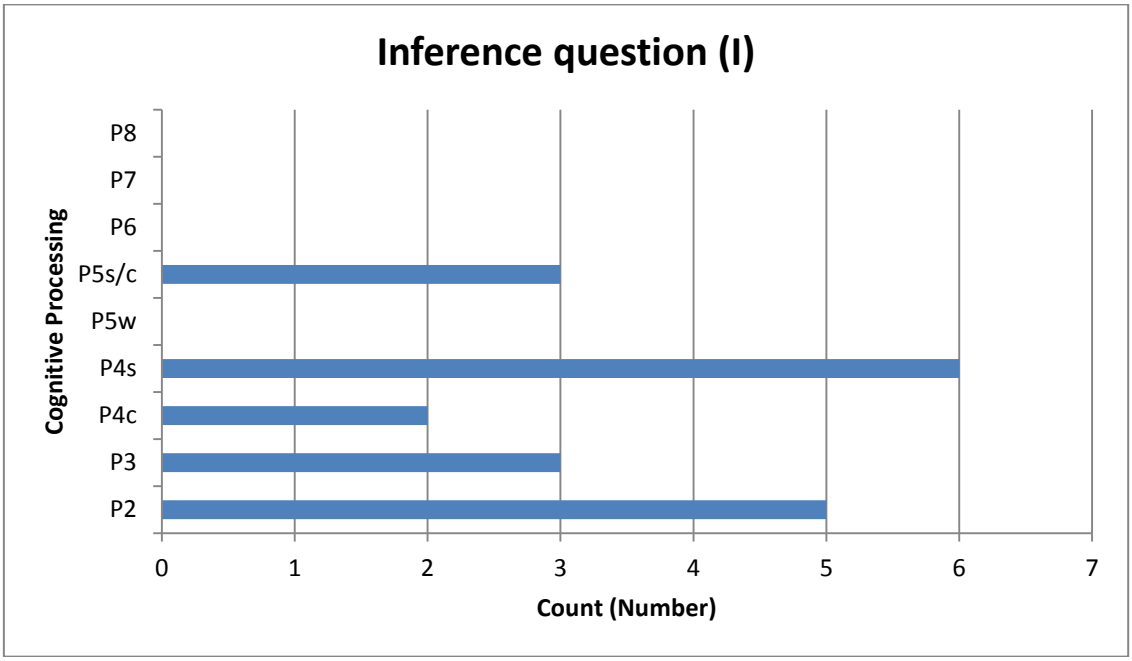


Figure 4.15. Frequency of cognitive processes for TOEFL inferencing questions

	P2	P3	P4c	P4s	P5w	P5s/c	P6	P7	P8
Total Processes	**5**	**3**	**2**	**6**	**0**	**3**	**0**	**0**	**0**
Processing ratio	**2.50**	**1.50**	**1.00**	**3.00**	**0.00**	**1.50**	**0.00**	**0.00**	**0.00**

Table 4.42. Most frequently-identified cognitive processes for TOEFL inferencing questions

Inference question (I)

"measure examinees' ability to comprehend an argument or an idea that is strongly implied but not explicitly stated in the text" (ETS, 2003: 25 in Cohen and Upton, 2006: 74).

Sample items

Text reference: "The Whigs, in contrast viewed government power positively. They believed that it should be used to protect individual rights and public liberty… In particular, **Whigs in the northern sections of the United States also believed** that government power should be used to foster the moral welfare of the country. They were much more likely to favor social reform legislation and aid to education."

9. Which of the following can be inferred from paragraph 5 about variations in political beliefs within the Whig Party?

- They were focused on issues of public liberty.
- They caused some members to leave the Whig party.
- They were unimportant to most Whigs.
- They reflected regional interests.

Text reference: "As a general rule, **the higher a mountain is, the more recently it was formed**; for example, the high mountains of the **Himalayas are only about 50 million years old**. Lower mountains tend to be older… about **400 million years ago**, when the present-day continents of North America and Europe were joined, **the Caledonian mountain chain was the same size as the modern Himalayas**"

29. Which of the following can be inferred from paragraph 2 about the mountains of the Himalayas?

- Their current height is not an indication of their age.
- At present, they are much higher than the mountains of the Caledonian range.
- They were a uniform height about 400 million years ago.
- They are not as high as the Caledonian mountains were 400 million years ago.

Strategic actions by the participants were highly visible in relation to inference items. Seven individual strategies were observed at a 'very high' frequency rate. Participants read the stem carefully to identify the purpose of the question. They carefully read each of the options to attempt to understand propositional content of each, before moving to the text to identify the relevant parts of the text and read carefully around them. This was achieved by examining and comparing key noun phrases in both the options and the text. They frequently returned to the options to refresh their working memory of the content of specific options. Options were eliminated based on contradictions between them and the content of the text. When an option was chosen, participants returned to the text to ensure their understanding was consistent with their choice.

As there was a 'very high' frequency for seven strategies, this provided a lot of stimulus for discussion in the subsequent interviews. For four participants completing two items, nineteen individual processes were identified, resulting in two processes (P2; P4s) recording 'very high' frequencies and three recording 'high' frequencies (P3; P4c and P5s/c). The high proportion of 'lexical access' codings relates to the strategic actions of identifying key words in both the options and the text. Participants were then effective at explaining propositional meaning of the parts of the text that they successfully parsed. The item type elicited detailed explanations at the sentence rather than the clause level, an imperative for subsequent inferential reasoning. P3 was used when participants explicitly stated they were matching nouns. Verbalisations associated with P5s/c codings provided the basis for examining the nature of inferential reasoning relating to this item type. No verbalisations were coded as P5w, indicating that inference items attempt to elicit inferential reasoning by the participant beyond the use of linking pronouns across sentences.

In relation to item 9, participant 1 provided the following excerpts in relation to her actions:

> *"These two are the words ['variations' and 'political beliefs'] that I want to use to find the detailed information. In option 1, there is no other information [that] is useful. This is the obvious information in the passage, so it's not inferences, so I think this one of course [is] not the answer. At this moment, I read this sentence [2], I want to check if there are any information related to options 2 and 3 [item 8], but it seems not. Because from here is the answer of question 8. [Underlines 'northern sections'] It seems like the 'regional' [option 4]. I can't find any related information about option 2 and option 3, but option 4, there is some information, so maybe there are different attitudes of the government in the different parts of the US, so I think maybe it is."* [Participant 1, test 2, item 9]

The participant identifies key noun in the stem and options. She presents verbal evidence of identifying main noun phrases, but not that she has established propositional meaning by linking them to the question stem or any portion of the text at this point (syntactic parsing).

She states that this item "may be related to government" as the highlighted noun ('concept') is linked to the post-nominal modifier 'government' in a prepositional phrase (syntactic parsing). The participant eliminates option 1, stating that this option refers to "obvious information in the passage" which does not fit the parameter of the item (inference). Ten seconds earlier, the participant had underlined 'individual rights' in sentence 2. She was able to link the content of sentence 2 to the Whig Party (mentioned in sentence 1 via the pronoun 'they'), and this is a positive affirmation of public liberty as one of their core beliefs. This is evidence of the participant 'establishing propositional meaning at the sentence level' to eliminate option 1 in item 9. The participant returns to the text underlines 'northern sections' in line 6 of paragraph 5. She links this to the term 'regional' in option 4 (lexical access). She states that she is unable to locate information related to options two and three. Having already eliminated option 1, she selects option 4 (correct).

In relation to the same item, participant 4 initially adopted the same strategy of identifying key words and relating these to equivalents in the text:

> *"It's the main point of the question. It's about variations of the beliefs, so there I will find some variations parts in the context. I think choice number 2 and number 3, there is no context in this paragraph. The Whigs think the government power should do this… the duty of the government. The same reason, because they are all about the government. I'm thinking about the fourth point, 'they reflected regional interests'… um, I think there is nothing about regions, so I found maybe this one is not the choice. Because the contents says about public liberty [option 1], and the other three [options] is [are] not mentioned in this context."* [Participant 4, test 1, item 9]

She similarly eliminates options 2 and states that there is no specific context for these options, leading her to mentally eliminate them. This suggests that her strategy for accepting or eliminating options at this moment is based on *lexical access* – searching for key words. She then underlines two infinitive verb clauses, identifying these as summarising the key aims of the Whig Party (establishing propositional meaning at the sentence level). She returns to the item and turns her attention to the options. She dismisses the fourth

option on the basis that there is no reference to 'regional interests'. To answer item 9 correctly, the participant needs to integrate the content of sentences 4 and 5; identifying that there were political differences between Whigs in the north and south, and recognising that the pronoun 'they' in the final sentence represents not 'Whigs' in general, but Whigs in northern regions. Her selection of option 1 stems from underlining and focusing on line 2 ('individual rights and public liberty'); evidence only of lexical access.

Item 9 rests upon the participants identifying the new and given information in the bold section of the text reference; 'also' believed' presupposes that a statement about the beliefs of some members of the Whig party was included in a preceding sentence (sentence 1). The participant is required to infer that the information contained in the predicate is *new* rather than given, and relate this back to the noun phrase predicate ('the Whigs') and the argument (beliefs stated in addition to those already covered in previous propositions) in sentence 1, and would present a new argument containing different information in relation to the predicate 'Whigs believe'. This additional information is identified in the predicate as being associated with members of the Whig party in one geographical location ('in the northern states'). 'Also 'believed' implies that Whigs in northern states share the fundamental beliefs with Whigs in other states, but have an additional set of beliefs beside these which are not shared by Whigs in those other states.

Item 29 is the second inference item in the TOEFL test used in the study. This item is presented in the same format as item 9, with three distractors presenting propositions that cannot be inferred from the information presented in the text. Participants 3 and 6 completed this item. Both recorded correct responses, although participant 3 changed her answer twice and revisited the question after she had completed the others in this section. She spent thirty seconds parsing the stem and options before selecting option 1, then returns to the text for one minute before eliminating option 1 and selecting option 3. However, at the end of this portion of the test, at 11.42, she returns to this question, eliminates option 3 and finally selects option 2 (the correct response). She initially selected option 1 as she in the text she has gained the impression that the height of a mountain may be misleading regarding determining their age. Option 1 is incorrect, as it contradicts sentence 2 in paragraph 2. The weakness of the participant's approach at this point is that

she attempts to answer the item using knowledge of the text that she has gained from the first reading – she responds without specifically referring to paragraph 2:

> *"I sort of tentatively went for number 1, 'their current height is not an indication of their age'. Because I had the impression that one of the points in the article is that only by looking at the height of the mountain often gives you the wrong idea of the age"* [Participant 3, test 2, item 29].

After she selected option 1, she returned to the text to reconsider her choice. She parsed the second sentence of paragraph 2 and demonstrates that she understands that this sentence contradicts option 1, leading her to eliminate it (establishing propositional meaning at the sentence level). She then selected option 3 on the basis that the text states that the Caledonian mountains and the Himalayas were similar-sized 400-million years ago (establishing propositional meaning at the clause level) although this choice neglects the parameters of the task in two ways. First, this is an affirmative statement made in the text, so does not meet the task requirements of inferring an option; secondly, the question stem only refers to the Himalayas rather than a comparison of the Himalayas and the Caledonian mountain range. After completing the remaining items in the test, she returned to page 3 to reconsider item 29. Twelve seconds after returning to page 3, the participant made her final (third) choice; eliminating option 3 and selecting option 2 (the correct response). She states that she had a sudden realisation that her understanding of the question stem and the second paragraph did not justify selecting option 3:

> *"I suddenly realised… the statement was not saying that at some point, they were at the same size, it was saying that at a particular time, they were the same size, but 400 million years ago, the Himalayas weren't even formed yet"* [Participant 3, 03:01:03:27-40].

The second sentence is particularly interesting, as at suggests inferential reasoning at the sentence level by the participant. The idea that mountain height is positively correlated with age is not stated directly in the text, but is an inference that the participant made as she

read of mountains being considered the 'epitome of permanence', stated in the first sentence of the second paragraph, forming in her mind the idea that the physically larger a mountain is, the longer it has remained in situ. This suggests that she may have also brought her own understanding of the world to the text when initially completing the item.

Participant 6 proceeded with item 29 quite differently. He spends considerable time thinking about item 29, ultimately selecting option 2. He admits to uncertainty regarding option 3 (he cannot make sense of the phrase 'uniform height'). Nonetheless, he notes options 2 and 3 for further consideration. The participant states he had to read in depth to ascertain the relevant information about the Himalayas. The participant correctly infers at the sentence level to select option 2:

> *"as a general rule, the higher a mountain is, the more recently it was formed. So, because the age of the Himalayas is younger, so it's higher than the Caledonian range"* [Participant 6, test 1, item 29].

The participant displays a form of logical thinking which he has used to impose meaning on the text and the item. This may be displayed syllogistically, with the test taker integrating two concrete propositions containing information about key nouns in the text (A and B) to arrive at a conclusion (C) that relates these two nouns together:

a.) The higher a mountain is, the more recently it was formed
b.) The Himalayas are younger than the Caledonian mountain range
c.) Therefore, the Himalayas are higher than the Caledonian mountain range

Proposition A is contained within the second sentence. The information for proposition B is contained within two sentences which describe the ages of both the Himalayas and the Caledonian mountain ranges These proper nouns act as anaphora for 'mountain range' in proposition A. Proposition C is not explicitly stated at any point in the text, requiring the participant to infer it from the previous two propositions. This is an example of a 'scaffolded' inference, as it is the product of an item produced to the item specifications. The test designers have encouraged this form of inferencing in the design of this item. In

order to guide the participant to this level of processing, the distractors promote this thinking process by referring to the age of the mountains, encouraging the test taker to focus on this aspect of the text. Each distractor must be incorrect for a very clear reason that the test taker is able to rationalise. This item may also be characterised as a *bridging inference* (Singer, 2007: 346) in that the linking sentences to provide the final proposition C rely on anaphoric references of 'mountain', 'Himalayas' and 'Caledonian (mountain range)' specifying the cognitive pathway for the participant to reach this proposition. Each of the options containing an anaphoric reference back to the question stem provides further grounds for suggesting that anaphoric bridging inference is the main understanding of 'inferential reasoning' encoded in TOEFL inferencing items.

4.5.2.6. (Inferencing) Rhetorical purpose question (I-rp)

Strategy code	Strategy label	Freq. rate
5	Reads question stem(s) and/or option(s) carefully	**2.00**
2	Identifies the purpose of the question	**2.00**
18	Returns to the question for clarification: rereads question and/or options	**2.00**
12	Marks/notes key noun phrase in the text during careful reading	**1.67**
11	Careful local reading (text)	**1.67**
6	Marks/notes key noun phrase(s) in the questions stem or options	**1.33**
27	Compares question stem/option to a portion of the text	**1.33**
1	Reads question(s) before proceeding to the text	**1.00**
28	Checks/confirms/considers option choice after reading portion of text	**0.67**
25	Eliminates option(s) (contradicts key noun/verb phrase/nonsensical response)	**0.67**
26	Eliminates option (s) (no information found)	**0.67**
31	Checks/confirms response has met the parameters of the task	**0.67**
8	Marks/notes key adjective phrase(s) in the question stem or options	**0.33**
13	Marks/notes key verb phrase in the text during careful reading	**0.33**
32	Uses own topic knowledge to enhance understanding of text/questions	**0.33**
24	Identifies paraphrase within text or between text and item stem	**0.33**
14	Marks/notes key adjective phrase in the text during careful reading	**0.33**
10	Marks/notes key adverbial phrase(s) in the question stem or options	**0.33**
30	Hesitates while answering to reconsider choice	**0.33**
29	Guessing	**0.33**
22	Identifies content-based parallel between paragraphs	**0.33**

Table 4.43. Most commonly-used strategies for TOEFL rhetorical purpose questions

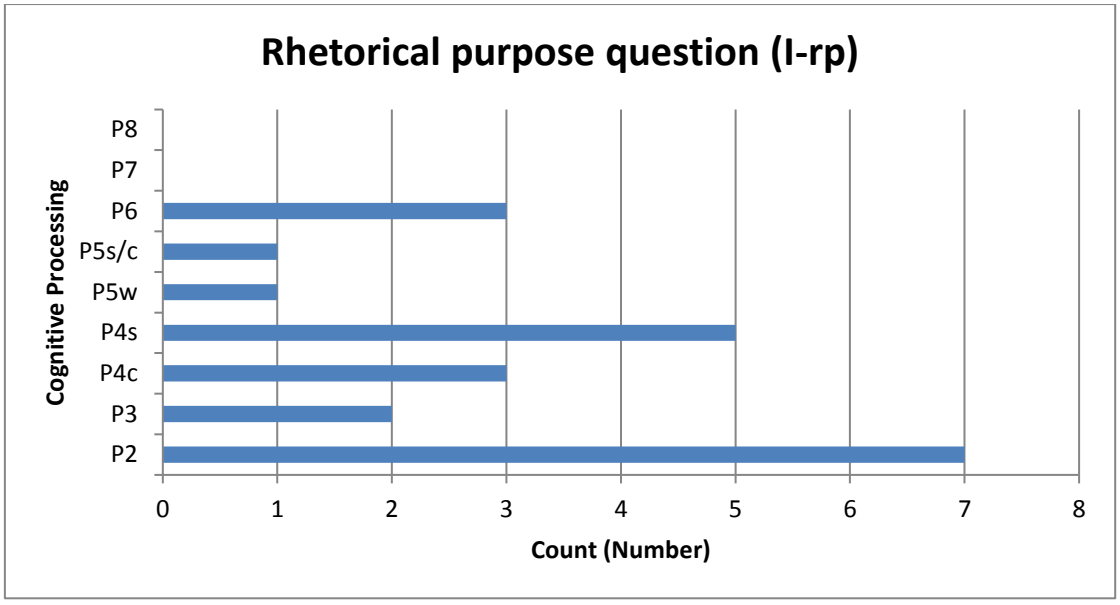


Figure 4.16. Frequency of cognitive processes for TOEFL rhetorical purpose questions

	P2	P3	P4c	P4s	P5w	P5s/c	P6	P7	P8
Total processes	**7**	**2**	**3**	**5**	**1**	**1**	**3**	**0**	**0**
Processing ratio	**2.33**	**0.67**	**1.00**	**1.67**	**0.33**	**0.33**	**1.00**	**0.00**	**0.00**

Table 4.44. Most frequently-identified cognitive processes for TOEFL rhetorical purpose questions

Rhetorical purpose question (I-rp)

"measure examinees' ability to identify the author's **underlying rhetorical purpose** in employing particular **expository features** in the passage and in **ordering** the exposition in a particular way. Correct responses require proficiency at **inferring** the nature of the **link between specific features of exposition** and **the author's rhetorical purpose**" (ETS, 2003: 27 in Cohen and Upton, 2006: 80).

Sample items

Text reference: "The Democrats tended to view society as a continuing conflict between "the people" —farmers, planters, and workers — and a set of greedy aristocrats. This "paper money aristocracy" of **bankers and investors** manipulated the banking system for their own profit, Democrats claimed…"

3. The author mentions **bankers and investors** in the passage as an example of which of the following?

- The Democratic Party's main source of support
- The people that Democrats claimed were unfairly becoming rich
- The people most interested in a return to a simple agrarian republic
- One of the groups in favor of Andrew Jackson's presidency

Text reference: "**Baring the teeth in a hostile way**, as noted by Charles Darwin in the nineteenth century, may be a universal sign of anger."

15. The author mentions **Baring the teeth in a hostile way** in order to

- differentiate one possible meaning of a particular facial expression from other meanings of it
- support Darwin's theory of evolution
- provide an example of a facial expression whose meaning is widely understood
- contrast a facial expression that is easily understood with other facial expressions

Text reference: "The weather, in its many forms, is the main agent of erosion. Rain washes away loose soil and penetrates cracks in the rocks. **Carbon dioxide** in the air reacts with the rainwater, forming a weak acid (carbonic acid) that may chemically attack the rocks."

32. Why does the author mention **Carbon dioxide** in the passage?

- To explain the origin of a chemical that can erode rocks
- To contrast carbon dioxide with carbonic acid
- To give an example of how rainwater penetrates soil
- To argue for the desirability of preventing erosion

Three examples of this item type were included in the test used in the research. Despite the few items available for analysis, a high frequency of strategy use was displayed for each. Three strategies recorded 'very high' usage, with a further five recording 'high' usage. Participants were careful to identify the purpose of the question by spending considerable time carefully reading the question, familiarising themselves with the parameters of the task and the relevant elements needed to successfully respond. Participants displayed evidence of frequently returning to the question to consider an option in relation to their developing understanding of the text. Despite the aim of the question to elicit higher level cognitive processes, participants still used search strategies to link key words in the stem and options to different parts of the text, and relied heavily on establishing propositional meaning of the sentence containing the highlighted word. Evidence included marking key noun phrases in the stem and then the text, despite them being already highlighted in both, and then read carefully around these words in the text. They compared the options to portions of the text to determine whether they matched their understanding of the text that they had read. Option elimination was a moderately used strategy based on verbal reports, which matches visual evidence of how participants engaged with this item type, as there is clear evidence that they engaged with successive options in an attempt to respond correctly.

Participants displayed evidence of a range of processing in completing this item type. Lexical access (P2) was recorded as 'very high'; commensurate with evidence of participants' lexical identification strategies. Additionally, establishing propositional meaning at the clause and sentence level and establishing a mental model recorded 'high' usage, consistent with efforts to read carefully around specific noun phrases. Syntactic parsing (P3) recorded moderate usage, indicating that this item type was not specifically targeting grammatical knowledge. Inferential reasoning at the word and sentence level recorded 'low' usage, with only one instance of each being coded, a surprising finding given the purpose of this item type. However, establishing a mental model was reported by three of the six participants, indicating that these participants felt that they had to get a firm grasp of the purpose of the text before they were able to answer this item type confidently. Inference in this item type is based upon participants forming a progressive understanding of the narrative and forming links across sentences to understand how the author builds an argument.

Following the strategic actions of participants, low-level processing (lexical access) was coded very frequently. Participants tended to begin by identifying relevant noun phrases that they believe will aid item completion in addition to the highlighted noun phrase. For example, participant 1 offered the following explanation of her actions in response to item 3:

> *"[I'm] locating information [which will] help me to understand the main point of this option. Maybe 'favor' is the key point of this option [four] 'in favour' or 'opposite'. [I'm reading] selectively, not in-depth"* Participant 1, test 1, item 3].

Regarding item 3, both participants 1 and 4 responded correctly (option 2). Participant 1 underlined a portion of the text that was unfamiliar to her (*'paper money aristocracy'*) on the basis that it is "very close to the question"; it is the head noun, including the noun phrase 'bankers and investors' connected to the highlighted phrase via the genitive construction. This action was coded syntactic parsing, on the basis that the participant could clearly see that it was linked semantically in the *context* ('of'), but was unable to offer an

explanation of how based on *content*. For this item type, participants then tend to read around the words they have highlighted in the text (*"[I'm reading the paragraph] in-depth. [I'm reading] around this one [bankers and investors]"*, participant 4, test 2, item 3).

However, once she had identified the relevant components, there is evidence that she links information from different parts of the text to build a mental model to assist her answering this item. She briefly returns to the previous page to refresh her memory about which party President Jackson belonged to (paragraph 1, sentence 1, a multi-clause sentence):

> *"So reading the options, I remembered… I want to check whether Jackson's group was the Democrats or the Whigs [so referred back to paragraph 1; "Jackson swept to power in 1829 at the head of the Democratic Party"]. So Jackson is the president of this party."* [Participant 1, test 2, item 3]

She contrasted this with the opponents (the Whig Party). This is clear evidence of enriching the proposition in paragraph 2 with information from paragraph 1, and therefore evidence of establishing a mental model beyond the sentence level. Identifying 'bankers and investors' as opponents requires linking the phrase to the two previous sentences, which elaborates a conflict between 'the people' and 'greedy aristocrats'. Her verbalisations imply a complex cognitive operation, whereby the participant identified Jackson's party; identified the difference between the Whig Party and the Democrat Party ('different attitudes towards commerce'); identified relevant members of each group, 'people' and 'bankers'; and established that the latter would not support the Jackson. This led her to eliminate option 4:

> *"But this one is talking about his opponents [who] formed the other party, so of course they are not in favour [reference to option 4]. After reading the whole passage I can't find something about the main source of support [option 1], so it's not this. At this time, I found the most likely information ["desire for sudden, unearned wealth"], the same as option 2."* [Participant 1, test 2, item 3]

This operation contrasts with the elimination of option 1. She is unable to find information related to the main source of support, so eliminates it without citing a portion of the text. In selecting option 3, the participant cites 'the desire for sudden unearned wealth' and links this to option 2 (''unfairly becoming rich'). Selecting the correct option was a cognitively more straightforward process than eliminating option 4.

Participant 4 also focuses her attention on the part of the text immediately preceding the highlighted phrase (lines 4-6). She cites the word 'aristocracy', stating that she does not understand it but that it might be important in answering the item. Remedial strategies at this point included moving between the options and the text to relate particular parts to see if other portions of the text might provide clues. One minute later, she selects option 2. She is able to relate the pronoun at the beginning of the sixth sentence to 'Democrats' in the preceding sentence (inferencing at the word level). Arriving at option 2 necessitates understanding of the information contained in the sentence with the highlighted words; these nouns form the subject of the fourth sentence which contains ideas of greed and unearned wealth (coded 'establishing propositional meaning at the sentence level'):

> *"[I went to] the question first. It's about the noun, two nouns; the implication of these two nouns. I think, maybe the paragraph is about what his opponents have done. [I'm reading] around this one [bankers and investors]. I have read this word, and they wanted the wealth offered without the competitive, changing society. And, these words 'bankers and investors', he [Jackson] said that… indicates that this one [as above] gained his [their] own profits in a very unfair way, so I chose this number 2."* [Participant 4, test 1 item 3]

The second rhetorical purpose question is item 15. This was completed by participants 2 and 5. Both participants responded correctly (option 3), displaying evidence that local engagement is sufficient for completion of this item. Participant 2 begins by selecting the shortest option to consider first, but disregards it on the basis that it is not related to the content of the sentence containing the highlighted phrase. In selecting the correct option, the participant cites the second half of the sentence in which the bold phrase appears – 'a

universal sign of anger', which for her paraphrases option 3 'a facial expression which is widely understood'; evidence that the participant has established propositional meaning at the sentence level:

> *"I ticked [option] 3 because I think this option is most relevant to the context. I started with the easiest [option]; 'support Darwin's theory of evolution'. I don't think this is remotely related so I just removed it as a possible answer. I just simply think, I'm sure that the other three aren't relevant because that is what the author tries to say – option 3. I guess I also read the words which is very helpful, that is the sentence below the one in bold letters. A 'universal' sign of anger. It reflects… re-emphasises what the author tries to say, that is referenced to the option 'widely understood'. It's like a paraphrase [of] 'a universal sign of anger'."* [Participant 2, test 2, item 15].

Participant 5 also proceeds by examining each option in turn. After briefly skimming the paragraph, she returns to the item and selects options 1 and 2 for consideration. She admits that she had not read the sentence before the highlighted phrase and the sentence containing the phrase in enough depth to establish propositional meaning. Her actions are initially based on surface-level understanding of the text. She returned to the text to read the section more carefully after reading the question and options. She then eliminates option 1. She notes that the word 'universal' gives the sentence meaning; facial expressions have 'universal value' (evidence of 'establishing propositional meaning at the clause level'). This helps her to narrow down the correct answer to option 2. She also selects this on the basis of local engagement with the text:

> *"I find out that I missed out this easy, other point. I think in this theory, it's actually trying to say that the facial expressions have this kind of universal meaning, so I re-underlined this, then it helps me to think, OK, so the right answer should be the second"* [Participant 5, test 1, item 15].

The third example of this item type is item 32, completed by participants 3 and 6. It was answered incorrectly by participant 3 (who selected option 3) and correctly by participant 6 (option 1). Participant 3 begins by reading the question stem and options, initially considering option 3 – she then selects it sixteen seconds later. She states that after viewing the options, she briefly moved to the text to read the highlighted phrase in context before returning to the question to select option 3. She considered option 1 "as there is truth in it", although she does not relate this statement to a specific part of the text, suggesting that she used schematic knowledge to make this statement. Option 3 is selected because the participant correctly links the sentence containing the highlighted term with the previous sentence, thereby enriching the proposition containing the highlighted phrase ('it's a continuation of the idea') and building a mental model of the paragraph:

> *"Again, this time I went directly to the question, and then after I read all the choices, I went back to the article again just to confirm, because after I read the choices, the first one, [has] ambiguity in it… 'carbon dioxide' is a kind of chemical, and it can erode rocks, but does the article talk about the origin of this chemical? I don't recall the article talking about that. But again, there is part of [the] truth in it, so I think I gave it tentative consideration, and then to contrast carbon dioxide with 'carbonic acid', I read the sentence again when it talks about carbon dioxide, and certainly not comparing the two of them, because carbonic acid is simply the outcome after the interaction with the chemical and the rocks. Um, 'to argue for the desirability of [preventing erosion]', to give an example of how rain water penetrates the soil… this was really my more favourable choice that I wanted to go for, but then I read the sentence again, just wanted to make sure that what does it mean to say that rainwater penetrates soil? Because the previous sentence talks about 'rain washes soil and penetrates cracks', so is this sentence still a continuation of that, or… thinking about whether it's a continuation of that topic or it's an initiation for a new idea, so I had to read the sentence again, and then I decided it's a continuation of the idea, so I sort of confirmed with myself the choice"* [Participant 3, test 2, item 32].

The participant correctly recognises that carbon dioxide is itself a chemical, although the text does not explain the origin of carbon dioxide. This explains her claim of ambiguity in the item. She actually provides a verbal explanation that proves her processing is sufficient to select the correct response ("*carbonic acid is simply the outcome after the interaction with the chemical and the rocks"*), suggesting that she understands that carbon dioxide is the basis for carbonic acid. Therefore, the participant did not respond incorrectly due to inadequate or inappropriate cognitive processing, but rather because she used her schematic knowledge to tell her what she thought to be the most appropriate response.

Participant 6 began by reading paragraph 5 and underlining the object noun phrase in the first sentence; he cites this as this may be significant in anticipating the content of the questions (lexical access). He then continues reading carefully for approximately one minute before moving to read the stem of question 32. He notes the target words ('carbon dioxide') which he immediately located in the text (lexical access). He verbally summarises the importance of this phrase in the second sentence (establishing propositional meaning at the sentence level). He stated that he read the remaining options, but eliminated them, as they were either not mentioned (option 2), or in the case of option 3, is eliminated as he is uncertain of the meaning:

> *"I read the highlighted 'carbon dioxide' in the paragraph, and then I read from here, it was talking about the carbon dioxide reacts with water, rain water, producing acid, to erode the rocks, so I think this option is talking about the chemical effects, so I chose this one to explain. [I read other options], but 'to contrast carbon dioxide with carbonic acid', I don't think the paragraph mentioned it. And then, actually, I'm not sure what 'penetrates soil' means, what does it mean, I think this one, the first option is the most correct, has the most relation[ship] to the paragraph, so I chose the first one"* [Participant 6, test 1, item 32].

4.5.2.7. (Inferencing) Insert text question (I-it)

Strategy code	Strategy label	Freq. rate
6	Marks/notes key noun phrase(s) in the questions stem or options	3.00
5	Reads question stem(s) and/or option(s) carefully	2.67
2	Identifies the purpose of the question	1.67
1	Reads question(s) before proceeding to the text	1.67
28	Checks/confirms/considers option choice after reading portion of text	1.67
11	Careful local reading (text)	1.00
25	Eliminates option(s) (contradicts key noun/verb phrase/nonsensical response)	1.00
31	Checks/confirms response has met the parameters of the task	0.67
18	Returns to the question for clarification: rereads question and/or options	0.33
27	Compares question stem/option to a portion of the text	0.33
8	Marks/notes key adjective phrase(s) in the question stem or options	0.33
13	Marks/notes key verb phrase in the text during careful reading	0.33
24	Identifies paraphrase within text or between text and item stem	0.33
14	Marks/notes key adjective phrase in the text during careful reading	0.33
30	Hesitates while answering to reconsider choice	0.33
33	Writes note/labels part of text/item	0.33
9	Marks/notes key prepositional phrase(s) in the question stem or options	0.33
4	Identifies grammatical or content-based parallel between items/options	0.33
23	Identifies lexical parallel between parts of the text (matches words across paragraphs or heading)	0.33

Table 4.45. Most commonly-used strategies for TOEFL insert text questions

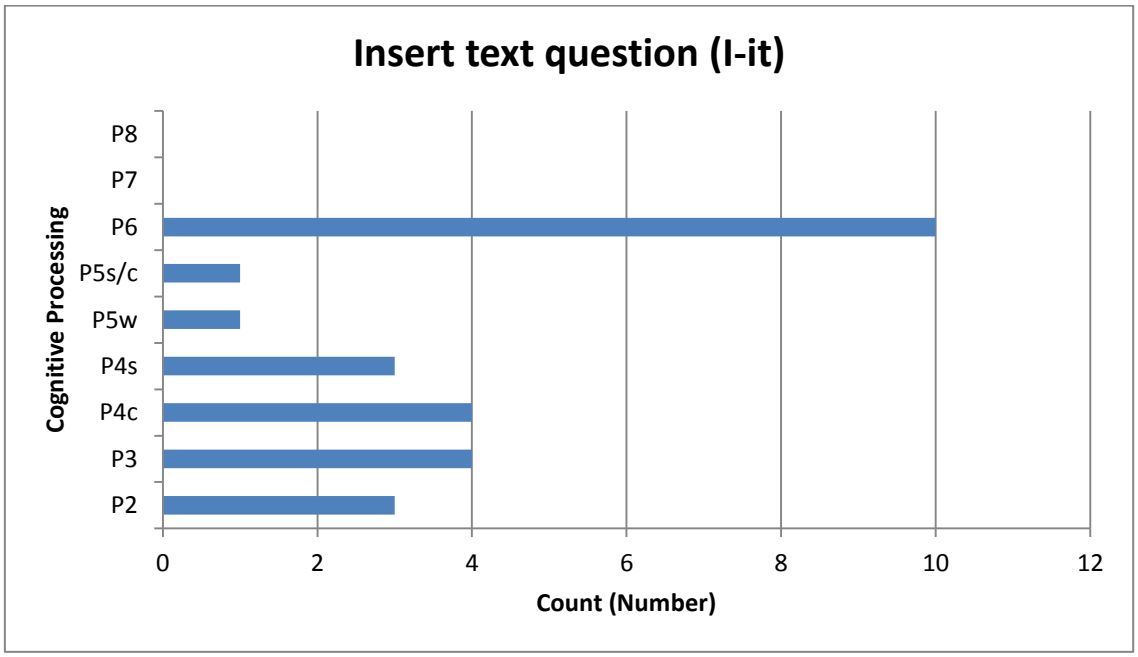


Figure 4.17. Frequency of cognitive processes for TOEFL insert text questions

	P2	P3	P4c	P4s	P5w	P5s/c	P6	P7	P8
Total processes	**3**	**4**	**4**	**3**	**1**	**1**	**10**	**0**	**0**
Processing ratio	**1.00**	**1.33**	**1.33**	**1.00**	**0.33**	**0.33**	**3.33**	**0.00**	**0.00**

Table 4.46. Most frequently-identified cognitive processes for TOEFL insert text questions

Sample items

Item 12

Paragraph 2. During Jackson's second term, his opponents had gradually come together to form the Whig party. ■ Whigs and Democrats held different attitudes toward the changes brought about by the market banks, and commerce. ■ The Democrats tended to view society as a continuing conflict between "the people" —farmers, planters, and workers — and a set of greedy aristocrats. ■ This "paper money aristocracy" of bankers and investors manipulated the banking system for their own profit, Democrats claimed, and sapped the nation's virtue by encouraging speculation and the desire for sudden, unearned wealth. ■ The Democrats wanted the rewards of the market without sacrificing the features of a simple agrarian republic. They wanted the wealth that the market offered without the competitive, changing society; the complex dealing; the dominance of urban centers: and the loss of independence that came with it.

Look at the four squares (■) that indicate where the following sentence can be added to the passage.

This new party argued against the policies of Jackson and his party in a number of important areas, beginning with the economy.

Where would the sentence best fit?

Item 25

Paragraph 2. ■ Most investigators concur that certain facial expressions suggest the same emotions in all people. ■ Moreover, people in diverse cultures recognize the emotions manifested by the facial expressions. ■ In classic research Paul Ekman took photographs of people exhibiting the emotions of anger disgust, fear, happiness, and sadness. ■ He then asked people around the world to indicate what emotions were being depicted in them. Those queried ranged from European college students to members of the Fore, a tribe that dwells in the New Guinea highlands. All groups, including the Fore, who had almost no contact with Western culture, agreed on the portrayed emotions. The Fore also displayed familiar facial expressions when asked how they would respond if they were the characters in stories that called for basic emotional responses. Ekman and his colleagues more recently obtained similar results in a study of ten cultures in which participants were permitted to report that multiple emotions were shown by facial expressions.

The participants generally agreed on which two emotions were being shown end which emotion was more intense.

Look at the four squares (■) that indicate where the following sentence could be added to the passage.

This universality in the recognition of emotions was demonstrated by using rather simple methods.

Where would the sentence best fit?

Item 37

Paragraph 6. Under very cold conditions, rocks can be shattered by ice and frost. Glaciers may form in permanently cold areas, and these slowly moving masses of ice cut out valleys, carrying with them huge quantities of eroded rock debris. ■ In dry areas the wind is the principal agent of erosion. ■ It carries fine particles of sand, which bombard exposed rock surfaces, thereby wearing them into yet more sand. ■ Even living things contribute to the formation of landscapes. ■ Tree roots force their way into cracks in rocks and, in so doing, speed their splitting. In contrast, the roots of grasses and other small plants may help to hold loose soil fragments together, thereby helping to pre vent erosion by the wind.

Look at the four squares (■) that indicate where the following sentence could be added to the passage.

Under different climatic conditions, another type of destructive force contributes to erosion.

Where would the sentence best fit?

There are three examples of this item type in the test used for this study (reproduced above), one relating to each of the three texts. Participants are presented with a paragraph from the text, with four positions in the paragraph highlighted with a square. Squares are located at sentence boundaries. Test takers are presented with one sentence written in bold and are asked which of the four locations the sentence best fits. Textual coherence should be established based on content, although cohesive text markers may also provide clues to the correct location. Clues may involve demonstratives such as 'this'; implying that the given sentence provides an example of a concept or idea stated in previous sentences. Linguistic clues may be cataphoric or anaphoric; that is, referencing parts of the text throughout the paragraph, both before and after the key location.

The definition of the item type offered by ETS suggests that a range of processing may be targeted by this item type:

> *"[insert text questions] measure examinees' ability to understand the lexical, grammatical, and logical links between successive sentences. Examinees are asked to determine where to insert a new sentence into a section of the reading that is displayed to them"* (ETS, 2003: 31 in Cohen and Upton, 2006: 86).

This definition suggests that participants will use lexical access (P2), syntactic parsing (P3) and establishing propositional meaning at the sentence level (P4s) as a minimum requirement. As the definition also states test takers will be measured on their ability to form *logical links*, this may also suggest the test developers intend this item type to target inferential reasoning (P5). As the definition targets *successive sentences*, this suggests that test takers are required to process locally rather than globally. This does not preclude forming a mental model, but suggests that test takers should not need to use information from other paragraphs to successfully complete this item type.

Data from participants identified two strategies that were very frequently used. These were 'marking key noun phrases in the stem/options' and 'reading the question stem and/or options carefully'. Strategies that recorded high frequency included 'identifying the purpose of the question', reading the question before proceeding to the text', checking an option after reading a portion of the text' and 'eliminating option (contradicts noun phrase)'. These strategies suggest some uniformity in how participants complete this item type. Participants tend to read the question stem and identify the purpose of the question. They read the target sentence carefully in order to identify key words which they use as clues to identify the most likely location of the sentence. They then read around the possible positions in the paragraph to see which parts are most likely to cohere with the sentence. Options could be eliminated on the basis that specific content does not coherer with the target sentence. The following discussion details how individual participants completed these items. Three items generated six individual accounts (two for each item).

Item 12 was completed by participants 1 and 4. Both participants respond correctly (option 1). Participant 1 began her exposition by directly comparing item 12 with the previous item

(11; sentence transformation). She states that the previous item was substantially more difficult than the insert text item ("*I found this type of question [11] … a little bit difficult. But this one, I think is not so difficult"*). The participant progresses to item 12. She then reads the instructions and focuses on the given sentence. She underlines the first three words ('this new party…'), identifying them as pronominal referencing that is influential in determining a coherent location for the sentence (syntactic parsing). In option 1, the participant highlights the 'Whig Party', leading her to consider option 1 on the basis of textual coherence (syntactic parsing). She goes on to consider option 2, and reads from the second sentence in the paragraph. She is evidently still looking at forming textual coherence based upon the opening pronominal phrase ('this new party'). She identifies detailed information in sentence 3 ('the Democrats tended…') which does not cohere with the content of the given sentence:

> *"I read option 2, I think there is no need for me to re-read the first sentence, and I go to the second sentence. I think this is the detailed information of the difference between Whig and Democrats, but this is to explain the attitudes of this new party, so I think it's not the correct order of the sentences in this passage. So I think it's wrong. This sentence seems like a summary of Jackson's attitudes, so maybe it should appear at the beginning of the paragraph"* [Participant 1, test 2, item 12].

This is evidence of building a mental model of the paragraph as the participant demonstrates understanding of both content and paragraph coherence. She understands that the sentence introduces ideas associated with the new party, so should appear towards the beginning of the paragraph. She looks at the second sentence in the paragraph, and summarises the content as being detailed information related to the new party, which should appear after the target sentence. The participant moves on to consider options 3 and 4, and states that she did not select these options for the same reason as option 2 – the surrounding sentences contained detailed information that did not cohere with the content of the given sentence – she is therefore further enriching the mental model she has built of the paragraph. She therefore selects option 1.

Participant 4 provides a very similar account of how she completed this item. She reads the instructions, briefly looks at the paragraph and then turns her attention to the target sentence. She also circles the pronoun 'this'; the first word in the target sentence. She states that the opening 'this new party…' will define what comes before; there will be a discussion of a new group (syntactic parsing). She reads the target sentence in the context of each option – she cites the 'opponents of Jackson' formed the Whig Party. She relates this directly to the phrase 'this new party at the beginning of the target sentence and therefore selects option 1 (evidence of building a mental model). She states that placing the sentence in this position makes the passage 'fluent'; evidence of her ability to integrate information across sentences:

> *"I'm trying to find which is the… the first option 'his opponents' come together to form the Whig Party'. They become the new party and the sentence put here, that makes 'this new party' accurate, so it makes sense. And it makes it more fluent I think"* [Participant 4, test 1, item 12].

There is therefore strong evidence that both participants completed this item successfully by integrating information across sentences and forming mental models of the paragraph. Participants explained that they understood the correct placement of the target sentence by eliminating those positions that did not cohere with the meaning of the paragraph – this led them to form an understanding of the progression of the argument in the paragraph, rather than relying on lexical cues across sentences, although these were important for forming initial impressions based on cohesive understanding of the text.

The second 'insert text' item is item 25, completed by participants 2 and 5. In contrast to participants 1 and 4, Participant 2 answers correctly (option 3) without displaying evidence of higher-level processing. Participant 5 also answers correctly, although displays a similar approach to this item as participants 1 and 4 for item 12. Participant 2 spends approximately one minute familiarising herself with the item and the sentence that needs to be inserted into the paragraph. This is followed by a flurry of action, as within 30 seconds, she draws a line in relevant part of the paragraph and underlines key words in both the text and sentence in the question stem (lexical access). The participant states that as she drew a

line between sentences (at position 3), she already considered this to be the correct response. She spends time underlining key noun phrases where she identifies synonymy between the sentence in the item stem and the second sentence in the text, indicating that she initially relied upon lexical access to respond to the item. However, she links 'universality' in the item stem to 'people in diverse cultures in the text', showing that she had, at minimum, 'established propositional meaning at the clause level'. There is no evidence that she processed at a higher level than this, by linking 'rather simple methods' in the stem to 'photographs' in the subsequent sentence. She answers this item quickly after the flurry of underlining, and actually states that she does not consider option four, presumably because she is confident of her choice:

> *"I spent some time on reading the question, when I drew the line, I already know the answer, that's the place this sentence should appear* [option 3]*. I underlined the words in the first sentence, 'same emotions' and 'all people'. You can see my mind is always chaotic, you can see I underline 'same emotions' and I underline 'recognition' and I underline 'recognise' and I underline 'universality'. I think they are synonymous between the same 'emotions' and 'all people', the two have universality. And 'recognition' and 'recognise'… that is the key words and the paraphrase correspondence between the words in the paragraph and the words you're working on from the sentence. [It's] Reassuring that 'recognise' and 'recognition'; this might be quite convincing that I might make the right choice. I didn't consider [position 4]"* [Participant 2, test 2, item 25].

In contrast to participant 2, participant 5 (responds correctly by selecting option 3) demonstrates evidence of building a mental model of the paragraph, aligning to the response patters of participants 1 and 4 in relation to item 12. Like those participants, he begins by reading the instructions and the given sentence. He circles the opening phrase ('this universality') then the gerundive phrase at the end of the sentence (lexical access). He spends approximately thirty seconds reading the paragraph, before circling position 3 in the paragraph. He moves to page 9 to see the sentence in context in option 3 before selecting it.

It is noteworthy that the participant also returns to briefly examine paragraph 1 in the event that the sentence is located in position 1 (at the beginning of the paragraph). This is an attempt to build a text-based mental model, although there is no further evidence that this strategy assisted the participant. Clues cited are the underlined parts of the sentence in bold ('universality' and 'simple methods') and that these relate to the sentences before and after the given sentence. The 'simple methods' refers to the subsequent experiment featuring photographs (building a mental model):

> *"I find out it was paragraph 2, then I read through it here, 'this universality…', then I thinking it's better I read the paragraph number 1, just in case if it's in the very front of this paragraph [in position 1]. I'm already quite sure about here, it should be option 3, so I read through it just in case [to make sure] it's logically correct. So, I find at the end of this sentence is 'using rather simple methods'. So I guess these words conclude what it says before this sentence, and this 'simple method' is what they are trying to say in the following part, so I think the third option is the best choice, because here, I understand that the first sentence and the second sentence, it [they] says about conclusions, and then in this sentence, it's already moved on to the experiments, 'taking photographs', so I think it should be put in here"* [Participant 5, test 1, item 25].

The third item of this type is item 37, the penultimate in the test. This was completed by participants 3 and 6. Both participants responded correctly (selecting option 1). Participant 3 commenced item 37 by reading the stem and summary sentence in bold. She spent approximately two and a half minutes considering the context of options 1 and 2 with the given sentence. It is noteworthy that she does not display any evidence of considering options 3 and 4, suggesting she was either able to dismiss them quickly or that she was highly confident in her initial choice.

When she turns her attention to item 37, she reads the sentence and uses her existing knowledge of the paragraph to consider likely locations – including at the beginning of the paragraph although she checks the parameters of the task and realises that this is not an

option. She then states that the sentence can only occur after 'ice and frost' because these are the initial climatic conditions under discussion (evidence that she had established propositional meaning at the clause level). She selects option one the basis of not only the previously stated climatic conditions ('ice and frost') but also because she notes the contrast with the subsequent sentence ('in dry areas'). This was coded as 'inferencing at the sentence level' as she contrasted the sentences based on content and inferred a shift in topic. She then cites the meaning of this sentence in relation to the subsequent sentence 'wind-driven sand' and that this is a different climatic condition – evidence that she is establishing a mental model of the paragraph. Although there is no evidence of her examining option 3 (on page 9), she does states that she examines this position without needing to see the sentence in context. She contrasts 'living things' in the sentence immediately after position 3 and states that this does not cohere with 'different climatic conditions' in the given sentence. This goes beyond lexical access and suggests that these different concepts do not cohere within the mental model she has built of the paragraph. She eliminates option 2 as it would involve separating a continuum of ideas across sentences – 'it' in sentence 4 relates back to 'wind' in sentence 3 (coded 'inferencing at the word level'):

> *"I think my attention moved back and forth between these two questions* [36 and 37] *at some point, but for this particular question, I did consider the possibility of whether this sentence could be in a different [position]… because the reason I read this is that I was thinking, 'this sentence can be at the beginning of this paragraph', but then of course I'd run back to see… and the square is not placed at the beginning of this paragraph. Then I considered, because this is also talking about climate condition, right? Then it can only be after it talks about 'ice and frost', because 'ice and frost' was [were] the initial climate conditions that were discussed, then it talks about another climate condition which is in dry areas, when it talks about wind-driven sand. And then in the third place, it's talking about living things, so that's not the climate conditions anymore, so I went for the first one. And then of course it cannot be in the second place because then that would be separating a continuous discussion of ideas; these two sentences cannot be*

separated, because they form a coherent idea because it's talking about wind" [Participant 3, test 2, item 37].

Participant 6 begins by reading the instructions and sentence in bold before turning his attention to the options showing the sentence in context. He cites the first prepositional phrase ('under different climatic condition') in the target sentence and states that he must find the point in the paragraph that talks about "another reason of erosion" (coded as 'establishing propositional meaning at the clause level'). Approximately one minute after reading the options, the participant selects option 1 (correct). He cites the opening sentence of the paragraph, noting it discusses 'cold conditions' (also coded 'establishing propositional meaning at the clause level'). Option 1 is cited as the most likely position as he cites the opening of sentence 2 ('in dry areas') which is a different climatic condition from cold conditions ('ice and frost'). This is consistent with the account offered by participant 3. He then states that position 3 is incorrect as 'living things' does not cohere with 'climatic conditions' in the target sentence and that this sentence coheres with the next sentence that discusses 'tree roots'. The participant here provides verbal justification that he uses logical progression of the paragraph to understand where the correct answer should be. He cites multiple sentences and how they connect to both explain why he selected option 1 and eliminated other options. As his discussion is at the paragraph level, this was used as evidence that he had successfully formed a mental model of this paragraph, subsuming the logical operations that are also clearly evident:

"I read the sentence that I need to put into [the paragraph]. I think it's 'under different climatic conditions, I think it must follow a sentence that has talked about… another reason of erosion. I read the sentence first, then found out it uses 'different climatic conditions', and another type is this word [cold conditions… ice and frost], and I think it must have a function of connection, like the paragraph talks about one of the reasons first, then this sentence followed the reason. Then another condition…, it's like in-depth [discussion] of it. I think option 1 is the most suitable for the logic format, is what I was thinking. I read the other options as well to make sure that the first option is correct. I just used the logic to think whether the format is correct. So it's

like, in dry areas, 'under different climatic conditions, another type of destructive forces contribute to erosion', but again, the following sentence is not talking about climatic conditions' [option 3]. 'Living things' is not 'climatic conditions'. It talks about 'tree roots' [as well]" [Participant 6, test 1, item 37].

4.5.2.8. Prose summary (R2L-ps) and Schematic table (R2L-st)

Strategy code	Strategy label	Freq. rate
6	Marks/notes key noun phrase(s) in the questions stem or options	**4.67**
12	Marks/notes key noun phrase in the text during careful reading	**4.33**
28	Checks/confirms/considers option choice after reading portion of text	**4.00**
11	Careful local reading (text)	**4.00**
33	Writes note/labels part of text/item	**2.33**
25	Eliminates option(s) (contradicts key noun/verb phrase/nonsensical response)	**2.00**
2	Identifies the purpose of the question	**1.67**
1	Reads question(s) before proceeding to the text	**1.67**
18	Returns to the question for clarification: rereads question and/or options	**1.67**
5	Reads question stem(s) and/or option(s) carefully	**1.33**
27	Compares question stem/option to a portion of the text	**1.00**
31	Checks/confirms response has met the parameters of the task	**0.67**
8	Marks/notes key adjective phrase(s) in the question stem or options	**0.67**
13	Marks/notes key verb phrase in the text during careful reading	**0.67**
30	Hesitates while answering to reconsider choice	**0.67**
26	Eliminates option (s) (no information found)	**0.67**
20	Skimming part of the text for general understanding (expeditious reading)	**0.67**
24	Identifies paraphrase within text or between text and item stem	**0.33**
14	Marks/notes key adjective phrase in the text during careful reading	**0.33**
10	Marks/notes key adverbial phrase(s) in the question stem or options	**0.33**
19	Searches for key word/phrase (text)	**0.33**
15	Marks/notes key prepositional phrase in the text during careful reading	**0.33**
16	Marks/notes key adverbial phrase in the text during careful reading	**0.33**

Table 4.47. Most commonly-used strategies for TOEFL prose summary and schematic table questions

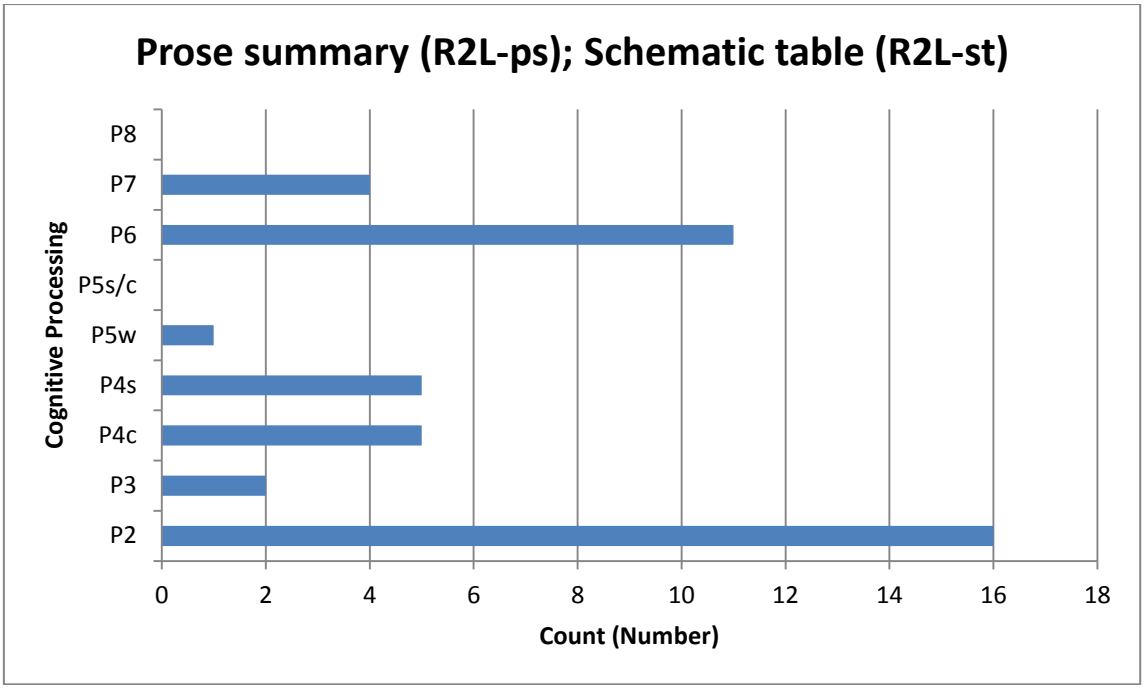


Figure 4.18. Frequency of cognitive processes for TOEFL prose summary and schematic table questions

	P2	P3	P4c	P4s	P5w	P5s/c	P6	P7	P8
Total processes	**16**	**2**	**5**	**5**	**1**	**0**	**11**	**4**	**0**
Processing ratio	**5.33**	**0.67**	**1.67**	**1.67**	**0.33**	**0.00**	**3.67**	**1.33**	**0.00**

Table 4.48. Most frequently-identified cognitive processes for TOEFL prose summary and schematic table questions

Sample items

Item 13

Directions: An introductory sentence for a brief summary of the passage is provided below. Complete the summary by selecting the THREE answer choices that express the most important ideas in the passage. Some answer choices do not belong in the summary because they express ideas that are not presented in the passage or are minor ideas in the passage. **This question is worth 2 points.**

The political system of the United States in the mid-nineteenth century was strongly influenced by the social and economic circumstances of the time.

1.
2.
3.

Answer choices

1. The Democratic and Whig Parties developed in response to the needs of competing economic and political constituencies.
2. During Andrew Jackson's two terms as President, he served as loader of both the Democratic end Whig Parties.
3. The Democratic Party primarily represented the interests of the market, banks, and commerce.
4. In contrast to the Democrats, the Whigs favored government aid for education.
5. A fundamental difference between Whigs and Democrats involved the importance of the market in society.
6. The role of government in the lives of the people was an important political distinction between the two parties.

Item 26

Directions: An introductory sentence for a brief summary of the passage is provided below, complete the summary by selecting the THREE answer choices that express the most important ideas in the passage. Some sentences do not belong in the summary because they express ideas that are not presented in the passage or are minor ideas in the passage. **This question is worth 2 points.**

Psychological research seems to confirm that people associate particular facial expressions with the same emotions across cultures.

1.
2.
3.

Answer Choices

1. Artificially producing the Duchenne smile can cause a person to have pleasant feelings.
2. Facial expressions and emotional states interact with each other through a variety of feedback mechanisms.
3. People commonly believe that they can control their facial expressions so that their true emotions remain hidden.
4. A person's facial expression may reflect the person's emotional state.
5. Ekmen argued that the ability to accurately recognize the emotional content of facial expressions was valuable for human beings.
6. Facial expressions that occur as a result of an individual's emotional state may themselves feed back information that influences the person's emotions.

Item 38

Directions: Three of the answer choices below are used in the passage to illustrate constructive processes, and two are used to illustrate destructive processes. Complete the table by matching appropriate answer choices to the processes they are used to illustrate. **This question is worth 3 points.**

Constructive processes	Destructive Processes
1.	1.

2. 3.	2.

1. Collision of Earth's crustal plates
2. Separation of continents
3. Wind-driven sand
4. Formation of grass roots in soil
5. Earthquakes
6. Volcanic activity
7. Weather processes

Prose summary and schematic tables are examples of ETS 'reading-to-learn' items. In both item types, participants are required to provide a summary of the text by selecting sentences from a list that best represent the overall argument contained within the text. No target paragraph is given for this item type. The participant is required to engage with the whole text to select the responses. This question appears last on the TOEFL paper, the intention being that the participant will use their knowledge of the text established in their working memory while answering previous items. The two reading-to-learn item types are analysed in a single section in the current study as the purpose is to identify and summarise the cognitive processes that ETS associate with 'reading-to-learn'. It is significant that ETS defines each of these item types separately, providing immediate clues regarding their intentions for each of these item types. ETS defines the summary completion item type as follows:

> "measure [ing] examinees' ability to understand the major ideas and relative importance of information in a text… An introductory sentence is provided, and examinees select 3 additional sentences from 6 options… [The three correct options] represent the major ideas in the text that, taken together, form a high-level summary of the text" (ETS, 2003: 15 in Cohen and Upton, 2006: 92).

Whereas the schematic table is defined as:

> "measure [ing] examinees' ability to conceptualize and organize major ideas and other important information from across the text…Correctly completed

> formats of these types reflect an able reader's mental framework of the text" (ETS, 2003: 18 in Cohen and Upton, 2006: 98).

The principle unit of analysis that this item type targets is the *text.* The aim is therefore to require participants to link information across parts of the text in order to form a coherent mental picture (P6 and P7 in Khalifa and Weir's model). There are several key differences between these definitions. Summary completion items require participants to *understand* the main ideas and their *relative importance.* This immediately indicates that distractors will be composed of minor ideas from the text that test takers must contrast with the main ideas, which form the keys. Conversely, schematic tables require participants to *conceptualise* and *organise* the major ideas in the text. Participants are therefore required to discern major from minor ideas in a text, but in this item type, they then must form a mental understanding of how these ideas relate to one another to create the overall argument contained within the text. The six participants each completed one of the three examples of this item the in the study. Participants 1 and 4 completed item 13, participants 2 and 5 completed item 26, and participants 3 and 6 completed item 38. Items 13 and 26 require participants to select three options from 6, with two marks awarded. One mark is awarded if participants select two correct options and no marks awarded if participants select one or no correct options. Item 38 requires participants to select 5 options from 7, but then to also categorise them correctly. Three marks are available. Participants receive three marks if all five choices are correct. Four correct choices receive two marks and three correct choices receives one mark. In the context of this item type, 'correct' means that both the options must be correct and be placed in the correct column. A correct option that is placed in the incorrect column is marked as incorrect (ETS, 2009).

Participants scored well in reading-to-learn items. Participants 1 and 4 both received full marks for item 13 (correct options 1, 5 and 6). Participant 5 completed item 6 correctly, receiving two marks by selecting options 2, 4 and 6). Participant 2 scored one mark (selecting options 2, 5 and 6). Participant 3 completed item 38 successfully, receiving full marks (options 1, 5, 6; 3 and 7). Participant 6 was awarded two marks (selecting options 1, 5, 6; 3 and 4). Thus the data provides complete records of accurate responses and

verbalisations associated with incorrect responses with which to contrast them. Verbal records were obtained for each of the choices that the participants made.

A wide range of observable strategies were employed for these items. Six strategies were rated 'very high' in terms of usage. These were 'marking key noun phrase(s) in the questions stem or options'; 'marking key noun phrase in the text during careful reading'; 'checking, confirming or considering option choice after reading a portion of text'; 'careful local reading (text)'; writing a note or labelling part of text/item' and 'eliminating options' because they contradict or otherwise do not cohere with part of the text that the test taker is focusing on. Despite the intention of the developers to target higher level processing, these strategies which are visible throughout item completion for 'reading-to-learn' items suggest that test takers will still use lower level processing to access parts of the text and orient themselves before engaging in more depth. Participants verbalise that this is their intention. Participant 1 for item 13 states:

> *"This is the two main nouns in this option. I'm looking for the… at the beginning, I think in these six options, there are two kinds of information, one kind is the comparison of the differences [between the Democratic and Whig parties] and one kind refers to some detailed information. Some information for this party [Democratic] and some information for the other party [Whig]."* [Participant 1, test 2, item 13].

The participant reads the instructions and the introductory sentence before moving to the list of options. She spends just over one minute reading the options and identifying key words, primarily noun phrases (coded as syntactic parsing because of the parts of speech the participant consistently identified). She also used key words to locate relevant information in parts of the text ("*I found 'education'"* [test 2, item 13]), clearly making a link between paragraph 5 and option 4 on the basis of the lexical item 'education'. Participant 4 adopts a similar strategy for this option (*"I'm trying to find the Whigs, and to see if they refer to the education… at that time, I didn't find something about the education"*) by searching the text for mentions of 'education' in relation to the Whig Party. She searches paragraphs 1 and 3, although does not find explicit reference to this option and for that reason,

eliminates it. Elimination was a common means of item management which provided significant evidence for how participants engaged with the text and options. In this instance, the participant focused on a key term and when she did not find it in the text, disregarded that option from further consideration. In contrast, the same participant eliminates option 3 and selects options 1 and 5 almost simultaneously. By way of explanation, she cites the noun phrases in option 3 ('markets, banks and commerce'):

> *"It's that, 'primarily represent the interests of the market, bank and commerce', I remember that in the passage, it's also talking about the society. And I still didn't find the specific position of this information, so I think it's not [this option] and it's also not the comparison of these two parties, just information about the Democratic [party]"* [Participant 1, test 2, item 13].

She uses her working memory to state that the text is concerned with society as a whole, not just these three areas, so option 3 represents detailed information, which is evidence that verbal explanations of eliminating options can provide evidence of high level processing; this explanation provides evidence that she has established a mental model of paragraph 2.

Reading-to-learn items provided substantial and clear evidence of higher level processing in the form of participant verbalisations. Two components of the cognitive processing model were rated 'very high' based on verbal evidence; 'lexical access' (P2) and 'establishing a mental model' (P6) of part of the text. Then high proportion of participant activity geared towards identifying key terms explains the high proportion of 'lexical access' coding'. The significance of forming a mental model of parts of the text is best exemplified by participants 3 (item 38) and 5 (item 26). Participant 2 scored 1/2 for item 26, as the evidence suggests that she failed to build a sufficient mental model of the text to select option 5 in place of option 4.

Participant 3 states that item 38 stimulated her working memory of how she approached the text in the initial six minutes of reading. She identified various constructive and

destructive processes which she highlighted as central to the progression of the argument made by the author before engaging with any items. Only eleven seconds after first reading the instructions to item 38, she selected option 1 ('collision of Earth's crustal plates') as a 'constructive process'. Seven seconds after this, she selected option 5 ('earthquakes') and one second later, option 6 ('volcanic activity'). She selected these three options without returning to the text: she cited them as constructive processes in one paragraph in particular (paragraph 3), from memory alone:

> *"I remember when I was reading the article at the very beginning, I noted down a few key words, and in one paragraph in particular, it talks about three different constructive processes, and I have impressions of them being crustal plates, earthquakes and volcanic activity, so that's why I very quickly picked out 1, 5 and 6"* [Participant 3, test 2, item 38].

Video evidence confirmed that she did not return to the text to select these options. Her explanation indicates that she has formed a mental model of paragraph 3. Once she has selected options 1, 5 and 6, she returns to the list of choices to consider those remaining. She spends longer considering the 'destructive' processes, eight seconds later selecting option 7; then seventeen seconds later selecting option 3:

> *"For number 7 I decided that very quickly because I remember one of the main destructive forces mentioned was weather processes. I hesitated about number 3 simply because I felt like 'wind-driven sand' was within the discussion of weather processes, so at that point, I was wondering whether there are maybe other options"* [participant 3, test 2, item 38].

The participant states that option 7 was easy to select as she remembered that 'weather processes' were described at length as a destructive process (paragraphs 5 and 6). She hesitates for longer before selecting option 3 as she states that she considered this option was subsumed by option 7. The ability to respond correctly to the item without reference to the text is evidence that she has command of the content of the text and how it cohered to create meaning.

Participant 5 quickly identifies the purpose and parameters of item 26. She moves to the options and immediately notes the 'Duchenne smile' refers to paragraph 5, which she marks in the option and on page 1 (lexical access). She writes '5' (indicating paragraph 5) next to option 1 before moving to consider option 2. Twenty seconds later, she selects option 2, without referring back to any portion of the text. She is able to state that this option provides an overview of paragraph 4:

> *"According to my general understanding of the whole article, according to my reading, when I was answering the following questions… I understand it's about facial expressions and emotional states, so I think it should be number 2, then I go back to paragraph 4 to check. It generally concludes the meaning of paragraph 4, I think"* [Participant 5, test 1, item 26].

She then turns her attention to option 3. She does not refer back to the text, but is able to state that this option does not think it is a clear representation of any portion of the text – the first clear indication by the participant that she is creating a text-level representation. She then considers option 4 and quickly selects it, stating that this option is a representation of paragraph 3. This option is a paraphrase of the opening sentence of the third paragraph. As she was able to select this option without referring back to the text, it was coded 'building a mental model' of this paragraph, showing her ability to identify the most important points. Regarding option 5, the participant notices the proper noun (the name 'Ekman') and remembers that he is cited in paragraph 5. This is coded as lexical access as the participant is only matching words, not ideas, and there is no evidence that she has selected the main idea in option 5 and related this to paragraphs 5 or 6 at this point. She then goes on to consider option 6. Approximately thirty seconds after reading this option, she selects it. She focuses on the verb ('influences') and relates this to the main purpose of the article:

> *"…because it says in the first part of the whole article, emotions produce expressions, basically, something like that, and in the second part of the article, it says like those facial expressions also can influence your*

emotions, maybe intensify it, or [make it] less intense" [Participant 5, test 1, item 26].

This is further evidence that the participant has created a text-level representation and that this option references paragraphs 5 and 6 in encapsulating the meaning of the 'facial-feedback hypothesis'. She then writes '2' next to the sentence in bold, stating that this sentence summarises the content of paragraph 2, successfully indicating that the participant has built a mental model of this paragraph also and has a genuine command of the text:

"According to the pattern, I think each sentence concludes a paragraph's meaning, so I guess I already got paragraph 2, paragraph 3, paragraph 4, I think then it's paragraph 2, then it makes sense. It almost covers the whole article's meaning. Because 'across cultures', so it basically means, we already discussed that like even those places, like 'the Fore', they have no connection to the Western culture, but still respond in the same way, so, yeah, they share the same emotions across cultures. You mentioned those wrong answers, probably about the details, so this 'Duchenne smile' [option 1], this people's name [option 5], these are details. They are not wrong, but they are just details" [Participant 5, test 1, item 26].

Participant 2 was the only one of the participants who did not score full marks in the reading-to-learn items. As she contemplates the options for item 26, she hesitates regarding options 2 and 4 specifically. She adopted a strategy of eliminating options that were incorrect alongside attempting to identify affirmative evidence for option choices:

"Although I ruled out the things I pretty much need to rule out, the things left, I'm not sure about. Especially number 2 and number 4. Those two options I'm not sure about. Then I… guess I'm reading this [the sentence in bold] and try to put them in order… I'm not sure about, not OK with number 2, I mentioned it 'a variety of feedback', 'certain patterns' [paragraph 3], it's not the way to say 'a variety of…', I'm not sure about number 2, also number 4, I think number 4 is like, the sentence is not harmful, it's neutral.

> *You can say it's the right one or not the right one, the wrong one… That means it might be right… it might be wrong, and it's hard to find the evidence for that sentence. [It's] very general"* [Participant 2, test 2, item 26].

She is uncertain of the positive evidence reinforcing option 2 ('a variety of feedback mechanisms'). The participant focuses on the syntactic marker 'a variety of' rather than 'feedback mechanisms' mentioned in the text, such as 'electrical activity in facial muscles'. The participant is correct to focus on this part of the text as evidence for option 2, but remains concerned about the phrase 'variety' [syntactic parsing]. She then states that sentence 4 is 'very general' and so is therefore difficult to find explicit textual support [establishing propositional meaning at the clause level]. She shows no evidence of identifying that this option is the major theme throughout paragraphs 3-6; she remains too pre-occupied with attempts to locate explicit textual support for each of the options she chooses. The participant is much clearer about the facial-feedback hypothesis discussed in paragraphs 3-6. She has successfully established the meaning of this hypothesis and so can link option 6 directly to this theory. Her understanding built up as she progressed through the test, evidenced by her engagement with items 21 and 24, and paragraphs 3 and 4 specifically. Her assuredness in selecting option 6 is the only evidence offered by the test that she has processed the text beyond the sentence level and has formed a mental model of this hypothesis. However, her elimination of options 3 and 5 were correct. Nonetheless, she ultimately selected option 5, despite previously arguing against it as she was unwilling to select option 4 – the clearest evidence that she did not create a cognitive representation of the whole text, explaining why she scored 1/3 for item 26.

4.5.3. Comparing cognitive processes by item type in IELTS and TOEFL

Each of the previous sections outlined the cognitive processes associated with each of the item types and the evidence base in the form of participant verbalisations. This section brings the findings together to directly address research question 4; *Are cognitive processes associated with specific item types? Do individual item types target specific processes or do they elicit a range of processes?* This section proceeds by collating the findings from sections

4.5.1 and 4.5.2 and presenting them in a cognitive specification matrix inspired by Buck (2001). This is useful to compare directly across the tests to determine the composition of the cognitive processes in each of the tests. These matrices also form the basis of a discussion comparing the findings in the study with existing literature regarding claimed cognitive processes in each of the item types.

Buck (2001: 109) recommends the production of a specification matrix as a visual and easily-understandable guide to test construction. This idea has been reproduced in this section with item types and cognitive processes. This way, the composition of the test, in terms of which processes are elicited by which item types, can be easily conveyed for consumption by interested stakeholders. This specification design can be used to define the construct (see literature review, section 2.3) in terms of the cognitive processes. The composition of the matrix is similar to that of the strategy matrix presented in table 4.4. A scale weighting applied to the strategies used by the participants was also applied to the identified cognitive processes in the test. The totals of each of the cognitive processes for each item type were divided by the number of examples of that item type. These were included in the tables of cognitive processes in each of the item sections (4.5.1 – 4.5.2). Labels are applied to the values commensurate with the labels attached to the strategies, as follows:

Very high (VH) frequency	≥ 2.00
High (H) frequency	≥ 1.00
Moderate (M) frequency	≥ 0.50
Low (L) frequency	≥ 0.30
Sporadic (S) frequency	≤ 0.29

Table 4.49. Scale weighting of frequency of cognitive processes divided by the number of items per item type (adapted from Cohen and Upton, 2006)

This data produced a cognitive specification matrix as outlined in table 4.50 below. Raw data for this matrix is displayed in tables 4.51 and 4.52. This table represents the outcome of the RE in terms of the cognitive processing identified in the two tests and therefore directly serves the test comparability research agenda:

IELTS	(1-MC)	(2-IDi)	(2-ID)	(3-IDw)	(4-MI)	(5-MH)	(8-SC)	(9-SuC)	Processing level
P2	VH	VH	H	H	VH	H	H	VH	LOW
P3	M	L	M	S	VH	S	M	M	
P4c	VH	M	H	M	H	S	M	M	
P4s	VH	H	M	M	H	H	M	M	
P5w	S	L	S	S	S	S	S	S	HIGH
P5s/c	S	S	S	S	S	L	S	S	
P6	H	S	S	M	L	H	S	M	
P7	S	S	S	S	L	S	S	S	
TOEFL	**(BC-v)**	**(BC-f/nf)**	**(BC-ss)**	**(BC-pr)**	**(I-rp)**	**(I)**	**(I-it)**	**(R2L-ps); (R2L-st)**	**Processing level**
P2	VH	VH	VH	M	VH	VH	H	VH	LOW
P3	M	M	H	H	M	H	H	M	
P4c	H	VH	VH	H	H	H	H	H	
P4s	L	H	VH	VH	H	VH	H	H	
P5w	L	S	S	S	L	S	L	L	HIGH
P5s/c	S	S	L	H	L	H	L	S	
P6	S	M	S	S	H	S	VH	VH	
P7	S	S	S	S	S	S	S	H	

Table 4.50. Weighted frequency of inferred cognitive processes for IELTS and TOEFL

Processing	1-MC	2-IDi	2-ID	3-IDw	4-MI	5-MH	8-SC	9-SuC	Total	Higher/lower level processing
P2	21	9	14	4	14	7	13	11	**93**	
P3	2	1	5	1	10	0	6	3	**28**	
P4c	12	2	8	2	8	1	4	3	**40**	**206**
P4s	10	3	7	3	9	6	5	2	**45**	**88.03%**
P5w	0	1	0	0	0	0	1	0	**2**	
P5s/c	0	0	0	1	1	2	0	0	**4**	
P6	5	0	1	2	2	6	2	2	**20**	**28**
P7	0	0	0	0	2	0	0	0	**2**	**11.97%**
Total	**50**	**16**	**35**	**13**	**46**	**22**	**31**	**21**	**234**	

Table 4.51. Profile of cognitive processes for IELTS (raw data)

Processing	BC-v	BC-f / BC-nf	BC-ss	BC-pr	I-rp	I	I-it	R2L-ps / R2L-st	Total	Higher/lower level processing
P2	20	48	8	1	7	5	3	16	**108**	
P3	5	9	3	3	2	3	4	2	**31**	
P4c	12	27	11	2	3	2	4	5	**66**	**256**
P4s	3	19	6	4	5	6	3	5	**51**	**82.58%**
P5w	3	1	0	0	1	0	1	1	**7**	
P5s/c	1	1	1	2	1	3	1	0	**10**	
P6	1	8	0	0	3	0	10	11	**33**	**54**
P7	0	0	0	0	0	0	0	4	**4**	**17.42%**
Total	**45**	**113**	**29**	**12**	**22**	**19**	**26**	**44**	**310**	

Table 4.52. Profile of cognitive processes for TOEFL iBT (raw data)

This data can be used as the basis of strong comparative claims between the reading sections of IELTS and TOEFL. Sections 4.5.3.1 – 4.5.3.4 highlight specific findings from the data, and relate them to existing literature. Section 4.5.3.1 compares the outcomes of the research to existing literature relating to cognitive processing in different item types IELTS, noting and exploring similarities and differences. Section 4.5.3.2 examines the disposition of high-level processing in IELTS. Section 4.5.3.3 compares the findings of the research for TOEFL with existing literature regarding cognitive processing in TOEFL, and section 4.5.3.4 looks at high-level processing in TOEFL. Overall comparative claims are summarised in section 4.5.3.5.

4.5.3.1. Comparison of emergent cognitive processes to existing literature for IELTS

As noted in section 4.4.3 and the literature review, there is a paucity of research related to the cognitive demands associated with each item type, due to the lack of significant revision to the reading section of IELTS since 1995. Weir et al (2009b) note that both the IELTS handbook (2005) and website do not contain information related to the construct. A new version of the handbook was produced in 2007, also did not contain cognitive information. According to the handbook, at least one text will contain “detailed logical argument” (IELTS, 2007: 7), suggesting the content of this text is likely to require higher-level processing to

formulate meaning, although there are no guidelines to suggest that this text is accompanied by specific item types that target higher-level cognitive processing.

The IELTS website nonetheless contains short descriptions of the skills that are associated with each item type. With reference to Khalifa and Weir's model of reading, it is possible to hypothesise the level (higher/lower) that each item type attempts to elicit and compare this information to the findings of the present study. For example, identifying information (True/False/Not given) items require test takers to 'recognise specific information'. Specific information is located within sentences. So the likely highest level of processing that this description refers to in Khalifa and Weir's processing core is 'establishing propositional meaning at the sentence level' (lower-level processing). In contrast, summary or diagram completion items may require test takers to identify 'understand details and/or the main ideas of a part of the text' ('building a mental model' or 'creating a text level representation') which requires higher level processing in order to integrate information across multiple sentences and paragraphs. The hypothesised level of processing for each item type included in the present study is included in table 4.51 below, alongside the description of the skills for that item type from the IELTS website and the outcomes of the present study based on information from table 4.50. If any higher-level process is rated 'medium' or higher in table 4.50, this is labelled 'higher and lower' in the right-hand column. If high level processes are only rated 'low' or 'sporadic', this is labelled 'lower' in the right-hand column. Note that only the item types included in the present study are analysed:

Item type	**Skills tested (IELTS, 2015)**	**Hypothesised level of processing (Khalifa and Weir, 2009)**	**Level of processing observed in present study**
Multiple choice	"Many different reading skills including: detailed understanding of specific points or general understanding of the main points of the text."	Higher and lower	Higher and lower
Identifying information (True/False/Not given)	"Ability to recognise specific information given in the text."	Lower	Lower

Identifying writer's views/claims (Yes/No/Not given)	"Ability to recognise opinions or ideas."	Lower	Higher and Lower
Matching information	"Ability to scan a text in order to find specific information. Unlike Task Type 5 (Matching headings), it focuses on specific information rather than the main idea. You may have to find: specific details, an example, reason, description, comparison, summary or explanation."	Lower	Lower
Matching headings	"Ability to identify the general topic of a paragraph (or section) and to recognise the difference between the main idea and a supporting idea."	Higher and lower	Higher and lower
Sentence completion	"Ability to find detail/specific information in a text."	Lower	Lower
Summary/note/table/flow chart completion	"Ability to understand details and/or the main ideas of a part of the text. When completing this type of question, you will need to think about the type of word(s) that will fit into a gap (for example, whether a noun is needed, or a verb, etc.)."	Higher and lower	Higher and lower

Table 4.53. IELTS reading item types, claimed abilities measured by each type and level of processing identified in the present study
http://www.ielts.org/test_takers_information/question_types/question_types_-_ac_reading.aspx

Table 4.53 demonstrates that the findings of this study provide good evidence for the claims of the IELTS test developers. There is a high-level of congruence between the claims that the developers make, the hypothesised levels of processing and the levels of processing that participants actually used in the present study. However, there were two specific discrepancies which require further elaboration. First, participants in the present study performed poorly on 'identifying information' items. As can be seen from table 4.50, they did not use high-level processes to complete these item types, and table 4.53 suggests that they should not require this level of processing. Nonetheless, sections 4.5.1.2 and 4.5.1.3 clarified that participants' failure in these items was due to an inability to process at a sufficiently high level, indicating that high level processing was required to be successful in

this item type. Specifically, analysis of item 12 in the IELTs test suggested that the developers intended for participants to possess knowledge of complex lexis and to integrate this lexical information across three sentences. Successful completion would have resulted in a verbalisation which would likely have been coded 'forming a mental model' (P6). The level of processing required in this specific item was far greater than the description of this item type suggests.

Incorrect responses can provide evidence that the test takers did not process at a sufficiently high level to correctly answer the question. This can be seen in analysis of items 19 and 23. Participant verbalisations demonstrate that item 19 targets cognitive processing at a higher level than 'syntactic parsing' and 'establishing propositional meaning at the clause level'. The question is designed to activate test takers' memory of several discrete points of information within paragraph 3, which must be combined in order to respond correctly ('false'). 'Nineteenth century studies of the nature of genius' is contained within sentence 1 of the third paragraph, while the 'uniqueness of a person's upbringing' is discussed in lines 10-15 of paragraph 3. However, an accurate verbalisation of a correct response is unavailable to state conclusively the level of processing required for item 19.

Item 23 is challenging due to the length of the noun phrase before the main verb:

> ***23.*** *The ease with which truly great ideas are accepted and taken for granted fails to lessen their significance.*

Item 23 begins with an extended noun phrase before a periphrastic verbal construction (main verb). The noun phrase itself contains a compound clause. The complex sentence structure of item 23 creates a cognitive load for item 23 much greater than that of items 21 or 22. Graesser et al (2004) argue that the number of words before the main verb of the main clause is an index of syntactic complexity, because it places a burden on the working memory of the test taker and require test takers to keep many words in their working memory before identifying the meaning of the main clause. Additionally, the head nouns in item 23 (ease, ideas, significance) are abstract rather than concrete. Contrasting evidence of this issue occurs when participant 5 reads items 21 and 22 concurrently before returning to

the text. The participant was clearly more confident of keeping the information of these two items in her working memory before consulting the relevant portion of the text.

This analysis is further backed up by the clear difference in table 4.53 between stated and required levels of processing for 'identifying writers' view or claim' items. Developers specify that this item type requires only low-level processing, although participants displayed evidence of needing to process at a high level to successfully complete them. Two instances of 'forming a mental model' (P6) and one of 'inferencing at the clause or sentence level' (P5s/c) were evident. As this item type constituted the final three items of the IELTS test, the participants may have been more likely to report their understanding of the text as part of the explanations of how they completed this item type. Overall, 'establishing propositional meaning at the clause and sentence levels', and 'establishing a mental model' of part of the text were the most frequently used codes for this item type, suggesting a need to integrate propositional information across sentences to successfully complete this item type. A recommendation for the developers would therefore be to revisit this item type and ensure that the design parameters of the question stems account for the breadth of the information base needed to successfully select the key, or to amend the claimed level of processing for this item type.

4.5.3.2. High-level processing in IELTS

The range of item types clearly indicates that IELTS developers adopt a componential approach to reading, and in terms of cognitive coverage, conceive of some overlap between the item types, as there is no uniform approach to the number of each item type in each version of the test. Individual texts may be matched with as few as two item types, and it is possible that a complete test (composed of three texts) will not contain the full coverage of item types used in IELTS. Test design indicated that the developers intend multiple item types to target high-level processing, and the data indicates that this is the case. Higher-level processing in the IELTS test is spread across 'multiple choice', 'identifying information' and 'identifying writer's views/claims' (discussed above), 'matching headings' and 'summary completion' item types.

‘Multiple-choice’ items are designed to elicit a range of cognitive processes, as item one asks participants to identify the main purpose of the text, whereas item three specifically asks test takers about a specific detail relating to two universities mentioned in the text. Evidence from this study suggests that multiple-choice items can be designed to access both higher and lower-level cognitive processes. For lower-level items, test takers should not engage with the entire text if they are to answer these items quickly and efficiently. For multiple-choice items targeting higher-level processing, test takers need to engage with multiple parts of the text to select the key and eliminate distractors.

‘Matching heading’ items are claimed to test the ability of participants “to identify the general topic of a paragraph (or section) and to recognise the difference between the main idea and a supporting idea” (IELTS, 2015). There is strong evidence from the engagement of the two participants that this item type tests a range of both higher and lower cognitive processes. Participants formed mental models of the paragraphs based on the propositional content. Answering confidently required participants to ensure that all parts of a prospective paragraph heading are addressed within the paragraph and that it encompasses the main point of that paragraph.

‘Summary completion’ items are claimed to test participants’ ability to understand details and/or the main ideas of a part of the text. This includes the class of word(s) that will fit into a gap (noun is needed, or a verb, etc.) (IELTS, 2015). Despite participants’ lack of success in this item type, sufficient evidence was presented to note that the item type targets both high and low level processes. Participants used key words to identify relevant parts of the text and lexical features of the summary and to identify answers. There were also clear instances in which participants needed to gather evidence from more than one sentence in order to answer confidently due to lexical similarity (items 8 and 10) and understanding the progression of the summary (items 9-10), which they failed to do on more than one occasion.

There is also further evidence from tables 4.50 – 4.52 that gap-fill items in general (IELTS items ‘sentence completion’ and ‘summary completion’) target local reading and low level processes rather than higher level processes. The level of processing associated with gap-fill

item types remains an area of debate in the literature. Alderson (1979, 1980) argues that cloze-type items are unable to elicit higher level processing. Although Alderson (2000: 208) also argues that cloze and gap-fill tasks measure different aspects of language proficiency due to differing item design (true cloze items eliminate words systematically e.g. every seventh, whereas gap-fill remove specific words to target specific processing), IELTS gap-fill items showed little evidence of eliciting higher level processes. Only two instances of higher-level processing were coded, indicating that this item type *is capable* of eliciting higher-level processing, although the emphasis is strongly on lower level processing on the basis of evidence in this study.

4.5.3.3. Comparison of emergent cognitive processes to existing literature for TOEFL iBT

Unlike IELTS, the TOEFL test has undergone a recent revision, moving from a paper-based to computer-based, and ultimately, internet-based test. Each transition was marked by fundamental reconsiderations of what the test was measuring. Chappelle et al (2008) produced a volume detailing the validity argument for the new TOEFL iBT, including the reading section, which provides a basis for analysing the findings from the current study. The mandate for the changes came from a series of white papers which emerged in the 1990s. Enright et al (2000) were responsible for developing new reading items based on a new conceptual framework that included communicative language competence (Bachman and Palmer, 1996). This framework identified *reading purpose* as the key determinant of task difficulty. Reading purpose is identified as one of the key extralinguistic features associated with language tasks (Jamieson et al, 2000). Different item types would have different reading purposes and this would represent a hierarchy of difficulty.

The reading framework includes three distinct reading purposes, with each individual item type relating to one of those purposes. These are included in the published blueprint for the reading component of the test. Jamieson et al (2008: 71) felt that a complexity scale associated with different item types would allow an interpretive score scale as different processes and task variables would account for differences between stronger and weaker test takers. Thus the framework is to be understood hierarchically (2008: 118), with

participants needing to master the skills on the left-hand side of the framework before progressing to those on the right. Different purposes of reading are therefore regarded as a variable that determines task difficulty. The finalised specifications for the TOEFL iBT defining the three reading purposes are outlined below in table 4.54:

Reading Claim	***Test taker can Understand English Language Texts in an Academic Environment***		
Sub-claims	***Basic comprehension:*** *can understand the lexical, syntactic and semantic content of the text and major ideas; can understand important sentence-level information; can connect information locally.*	***Inferencing:*** *can comprehend an argument or idea that is strongly implied but not explicitly stated in the text, identify the nature of the link between specific features of exposition and the author's rhetorical purpose, and understand the lexical, grammatical and logical links between successive sentences in a passage.*	***Reading to learn:*** *can connect information across the entire text; can recognise the organisation and purpose of a text, understand the relative importance/scope of ideas in a text; can understand rhetorical functions and purposes and organise (categorise/classify) important information into an appropriate mental framework representative of the organisation and inter-relationship of ideas in a text.*
Nature of reading task	*Questions about main ideas and supporting details based on individual propositions in a text, including vocabulary, reference, sentence simplification, factual information, or negative fact.*	*Questions about information or an idea that is implied but not stated, about the author's purpose in employing an expository feature, or about where in a text a new sentence should be inserted.*	*Questions that require test takers to create a summary of the main ideas and questions that require test takers to classify/categorise information into a schematic table.*
Response types	*Simple selected response*	*Simple selected response*	*Complex selected-response – prose summary, schematic table.*
Scoring rubric	*Dichotomous (right/wrong 0-1)*	*Dichotomous (right/wrong 0-1)*	*Dichotomous (right/wrong 0-1); particle credit (0-4)*
Number of questions	*9-28*	*9-18*	*3*
Nature of material	*Three passages on different topics selected from exposition, argumentation or historical background/ autobiographical narrative; all texts categorised as having one major focus of development or more than one focus (approximately 700 words).*		
Total time	*Approximately 60 minutes for 39-45 questions*		

Table 4.54. Final blueprint for TOEFL iBT Reading (Jamieson et al, 2008: 244)

The specification outlines the claims associated with the TOEFL iBT reading section. Three sub-claims were identified (Pearlman, 2008: 242) which relate specifically to each of the item types used in this study. These include an inferencing sub-claim that test takers are able to comprehend ideas that were not explicitly stated in a text, and that was a skill that could be distinguished from basic comprehension (ibid: 242). This final blueprint is reflected in the Official Guide which is presented to test takers and institutions. The 'nature of the task' section provides substantive information regarding the developers' conceptions of which test items elicit higher or lower-level processes. Basic comprehension items refer to "main ideas and supporting details based on individual propositions in a text" (lower-level), with remaining definitions for inferencing and reading to learn clearly targeting higher-level processes. Table 4.53 below includes the hypothesised level of processing based on Khalifa and Weir's framework (2009) and the observed levels of processing based on data collected for this study:

Item type	**Reading purpose (Jamieson et al, 2008: 244)**	**Hypothesised level of processing (Khalifa and Weir, 2009)**	**Level of processing observed in present study**
Vocabulary questions	Basic comprehension	Lower	Lower
(Negative) Factual information	Basic comprehension	Lower	Higher and lower
Sentence simplification	Basic comprehension	Lower	Lower
Pronoun reference	Basic comprehension	Lower	Higher and lower
Inference	Inferencing	Higher and lower	Higher and lower
Rhetorical purpose	Inferencing	Higher and lower	Higher and lower
Insert text	Inferencing	Higher and lower	Higher and lower
Prose summary and Schematic table	Reading to learn	Higher and lower	Higher and lower

Table 4.55. TOEFL iBT reading item types, claimed abilities measured by each type and level of processing identified in the present study

Table 4.55 demonstrates that the findings of this study provide good evidence for higher-level processing in the TOEFL test. It is immediately noticeable that basic comprehension item types designed by ETS to target lower-level processing can elicit higher-level processing by test takers. The data provides evidence of low-level processing for 'vocabulary' and

‘sentence simplification’ basic comprehension item types, although suggests that ‘factual information’ and ‘pronoun reference’ item types can target higher-level processing.

‘Vocabulary’ item types target lower-level processing. Despite this item type ostensibly targeting ‘lexical access’, participants still read around the highlighted word carefully; verbal evidence suggested that word matching based on the meaning of isolated words is insufficient to successfully complete this item type. This is backed up by verbal explanations offered by participants. Although ‘lexical access’ (P2) was the most frequently used processing code, the second most frequently used was ‘establishing propositional meaning at the clause level’ (P4c), indicating participants felt they needed to establish the meaning of the word in context to respond confidently. Participants who responded very quickly to this item type relied upon lexical access. Those who were less sure applied remedial actions – reading the key and distractors in context to ensure that they had selected the correct option.

‘Sentence simplification’ items also target lower-level processing. Participants’ verbal explanations revealed that the most common cognitive processes were ‘lexical access’ and ‘establishing propositional meaning at the clause and sentence levels’. Despite the apparent complexity of the item type, there was only a single instance of higher-level processing observed; an instance of ‘inferencing at the clause/sentence level’ was recorded when a participant used her own schematic knowledge to eliminate an option that was not feasible.

Evidence suggests that sentence simplification items are therefore a manifestation of processing competence operating exclusively at the sentence level. Processing competence is concerned with the cognition of individual lexical items as encountered during reading and how these individual units are transformed into a complete idea. This process contains both an assembly and an evaluative phase, operating dynamically in which individual lexical items contribute towards a developing understanding of the sentence which becomes fully formed once the complete sentence is parsed. Structures are then evaluated in terms of semantic and discourse appropriacy. The key and distractors must be parsed so that this item type cannot be successfully. Word matching or matching linguistic structures is not a viable strategy for this item type.

This item type requires a grammatically and conceptually complete highlighted sentence for their effective functioning. Isolating a single sentence and directing a test taker to it coerces the test taker to adopt a 'garden path' model (Frazier, 1987) of sentence comprehension. That is, information about the sentence (lexical, syntactic) is employed as each word is accessed from the test takers working memory as the sentence is parsed out of context. Individual lexical units (words) trigger the schematic knowledge of the test taker and the subsequent retrieval of the referents (in the case of noun phrases), and these are combined with verb phrases: the constituent parts create one grammatically complete, interpretable structure.

4.5.3.4. High-level processing in TOEFL

The blueprint of the TOEFL test (table 4.52) clearly demonstrates that each of the item types is intended to target one of the three reading purposes, with the lowest ('basic comprehension') targeting lower-level processing most explicitly, with 'inferencing' and 'reading to learn' targeting higher-level processing. However, data from this study demonstrates that 'basic comprehension' also covers higher-level processing, as defined by Khalifa and Weir (2009). Evidence of high-level processing in TOEFL was found in six of the item types (table 4.55). These were 'factual information', 'pronoun reference', 'inferencing', 'rhetorical purpose', 'insert text', and 'prose summary and schematic table'.

The rationale for 'factual information' item types is testing understanding of propositional information in the text at the clause or sentence level. However, participants 3 and 4 displayed higher-level processing, specifically 'building a mental model' (P6) of the text in relation to these items. Additionally, participant 4 responded to item 2 displaying evidence of 'building a mental model' of the relevant paragraph by integrating information from multiple sentences, suggesting that command of the relevant parts of the text plays a significant role in participants eliminating distractors. This item type requires that participants read and consider all options carefully in the context of relevant information, pushing participants towards higher-level processing by forming mental models of parts of the texts.

A variety of processing was inferred from participant verbalisations in the 'pronoun reference' item types. Across the four participants who completed this item type, twelve specific processes were inferred, the most common being establishing propositional meaning at the sentence level. Evidence of inferential reasoning occurred in relation to item 17 due to the necessity of linking 'photographs' to the verb 'depicted'. In order to respond successfully to this item type, the participants have to form textual coherence between pronouns or other anaphoric devices and their given referents. Item 17 requires participants to form this link across sentences, a form of 'anaphoric bridging inference' (Singer, 2007: 346-347). Three additional item types claim to elicit inferential reasoning, which provides the working definition of this mental process in the TOEFL test. These are 'inferencing', 'rhetorical purpose' items and 'insert text' items.

'Inferencing' item types appear to prioritise a form of inferential reasoning known as syllogistic reasoning, as identified in section 4.5.2.5. Two propositions can be identified from information given in the stem. Each of the options represents a possible conclusion to be drawn from the propositions, of which only one may be logically derived. This item may also be characterised as prioritising 'bridging inferencing' (Singer, 2007: 346) in that the linking sentences to provide a final proposition rely on a clearly elaborated cognitive pathway for the participant to reach the correct proposition. Many of the options in this item type contain anaphoric references back to the question stem providing further grounds for suggesting that 'anaphoric bridging inferencing' is the main understanding of 'inferential reasoning' encoded in TOEFL inferencing items.

Despite the aim of targeting inferential reasoning, 'rhetorical purposes' items recorded 'low' usage of inferential reasoning at the word and sentence level, with only one instance of each being coded, a surprising finding given the intended purpose of this item type. However, establishing a mental model was reported by three of the six participants, indicating that these participants felt that they had to get a firm grasp of the purpose of the text before they were able to answer this item type confidently. A similar picture emerged for 'insert text' items. The definition of this item type from section 4.5.2.7 states test takers will be measured on their ability to form *logical links*, this may also suggest the test

developers intend this item type to target inferential reasoning (P5). However, this item type recorded ten instances of 'building a mental model' (P6) and only one instance each of inferencing at the word and clause/sentence levels.

Inferencing in these item types is based upon participants forming a progressive understanding of the narrative and forming links across sentences to understand how the author builds an argument. This would provide consistency with inferencing items, but it calls into question at which point 'inferential reasoning' becomes 'building a mental model' in Khalifa and Weir's model. As syllogisms are a form of mental model, this form of cognitive processing calls into question whether 'building a mental model' should be classified as higher-level processing than 'inferencing' in Khalifa and Weir's model of reading. Item 29 (inferencing) clearly tests a participant's ability to construct a mental model that the writer intends the reader to form by integrating information across propositions before inferring the conclusion from the major and minor premises. As a result, building a mental model is a requirement of inferential reasoning. Inferential syllogistic reasoning necessitates the ability to discern between major and minor propositions. Portraying cognitive processing as strictly hierarchical is an artefact of conceiving of bottom-up and top-down processes and linking processing to linguistic units (letter/phoneme/word/sentence/paragraph/text), whereas the data in this study suggests that identification of specific high-level processes can be ambiguous.

Data from Cohen and Upton (2006) suggested that inferential reasoning was under-represented in the TOEFL test in terms of the definition of inferential reasoning offered by the test developers, which is reproduced here:

> "can comprehend an argument or idea that is strongly implied but not explicitly stated in the text, identify the nature of the link between specific features of exposition and the author's rhetorical purpose, and understand the lexical, grammatical and logical links between successive sentences in a passage" (Jamieson et al, 2008: 244).

Data from this study provides the necessary data which demonstrates that this form of inferential reasoning is indeed present in the TOEFL test, although it is centred in the 'inferencing' item types rather than the 'rhetorical purpose' item types as implied in the quotation above. 'Rhetorical purpose' items instead recorded more instances of 'building a mental model' (P6), suggesting that greater contextual cues are necessary in understanding why specific language is used than forming logical links across sentences. This distinction calls into question the distinction between inferential reasoning and forming a mental model in Khalifa and Weir's model and suggests that this division requires further clarification if a hierarchical approach to cognitive processing is to form the basis of future investigation into high-level cognitive processing.

The final item type to elicit high-level processing is 'prose summary and schematic tables', of which there is one item type per text. This is the 'reading to learn' item in the TOEFL specification (table 4.54). The rationale of the item is conceptualising or ordering major ideas in the text. This item type recorded eleven instances of 'forming a mental model' and four instances of 'building a text-level representation', suggesting that this item type does not require participants to have mastery of the complete text. For example, item 38 requires participants to identify and organise supporting points to the main argument of the text – that the Earth is a dynamic body. There is no evidence in participants' verbalisations that they required knowledge of the *rhetorical structure* (Khalifa and Weir, 2009: 53; Enright et al, 2000: 5-6) to successfully complete this item as no part of the question requires integration of constructive and destructive processes into a narrative argument. For this reason, the code 'creating a text level representation' (P7) was underused relative to 'building a mental model' (P6), reflecting the localised source of information for this item type.

Although the TOEFL is dominated by multiple choice, an item design which may initially appear limited in the range of cognitive processes that can be elicited. The 'vocabulary' item type appears to reflect ETS' legacy of focusing on language knowledge rather than communicative ability, as these items may be answered correctly without reference to the text if the participant is familiar with the lexis. However, evidence from the present study demonstrates that multiple-choice items allow for a variety of cognitive processes to be

elicited. Even in instances in which participants were confident of their performance in 'vocabulary' items on the basis of their vocabulary knowledge, a common strategy was to return to the text and mentally insert the chosen word in the context and read around it to ensure that the word functions in that context. This is reflected in the high proportion of coding of 'establishing propositional meaning at the clause level' for this item type.

4.5.3.5. Summary of comparative claims of strategy use and cognitive processing in IELTS and TOEFL

RE of IELTS and TOEFL revealed a number of similarities and differences between the reading components of IELTS and TOEFL identified through the SRI conducted with the six participants. Data reflects the general trend that strategy use was more prevalent amongst participants when they were completing the TOEFL test than the IELTS test. The data reflects the general observation that most strategies were used sporadically across all item types, although overall, participants recorded more intermittent engagement with the text in IELTS tests than in TOEFL tests. This is at least partially due to test method effect. In TOEFL, participants recorded more instances in which the participant returns to the question for clarification (monitoring/goal checking). Participants were more likely to read the questions before proceeding to the text in TOEFL than in IELTS due to the length and complexity of question stems. Fewer instances of observed strategizing in relation to the IELTS test led to a paucity of input in IELTS interviews compared to TOEFL. This led to higher frequencies of all cognitive processes in TOEFL compared to IELTS.

The generally higher level of strategizing observed in relation to the TOEFL test has resulted in greater proportions of processing being observed in TOEFL than IELTS.
For TOEFL, higher level processes are associated with items that concentrate in the second half of the test. Four basic comprehension item types contain significantly fewer higher level processes than the subsequent four item types (inferencing and reading-to-learn). TOEFL reading-to-learn items target the top end of the Khalifa and Weir (2009) model and require participants to form mental models of parts of the text. This item type provided strong evidence of these levels of the model, in contrast to IELTS, where evidence of participants'

text-level comprehension was less clearly associated with specific item types. For this reason, participants in the IELTS test may encounter an item type that requires processing at a high level as an opening question. For TOEFL, the majority of items participants encounter at the beginning of the text will target local reading relating to specific details in the text. Concentrating more cognitively demanding items in the second half of a test is one means of creating more discriminatory power in a test. Hypothetically, items that are moved from the beginning of the test to the end become more difficult as participants are increasingly time-constrained. The difficulty of an item may therefore be associated with the order of items in the test paper, independent of cognitive demand. Placing cognitively demanding items at the end of a test will advantage higher-level test takers. High ability test takers who have formed an overall impression of the text as they have progressed may find that they are able to respond quickly to items that test their global understanding of the text. Low ability test takers who have not established a mental model of the text will struggle with the cognitive demands and the lack of time available to complete the remaining items.

It is clear that item design can influence participants' engagement with the items sufficiently to influence their cognitive processing. More numerous instances of observable participant engagement with the TOEFL test may be attributed to the length of item stems and options in TOEFL, which are considerably longer than those in IELTS, resulting in greater visible engagement with the items themselves and a corresponding greater tendency to return to the item options for clarification. In contrast, item types such as 'matching headings' in IELTS contain only a single stem for multiple items, with option choices composed of only a few words, allowing participants to hold information from multiple options in their working memory. This contrasts with 'sentence simplification' in TOEFL, which required very careful engagement with each of the options to eliminate them and to answer confidently.

IELTS item types each target a range of cognitive processes. Evidence presented here provides justification for IELTS test developers designing tests which do not contain the full range of possible item types in each version. Individual texts may be matched with as few as two item types, and it is usual for a complete test (composed of three texts) to not contain the full range of item types used in IELTS. The flexibility with which IELTS developers include

different item types in different test versions means that item types overlap in terms of cognitive coverage.

In contrast, each TOEFL test will contain the full range of item types designed by the developers. TOEFL item types are designed to target specific reading purposes, which relate to the levels of processing required for successful item completion. The three purposes are 'basic comprehension, 'inferencing' and 'reading to learn'. The latter two reading purposes most explicitly target higher-level processing. In TOEFL, these item types are arranged in a specific order so that items eliciting higher-level cognitive processes are located in the second half of each test. Specific item types targeting inferential reasoning in TOEFL provide clear evidence of this level of cognitive processing in the TOEFL test.

Inferential reasoning at the sentence level in TOEFL occurred most prominently in 'inference' and 'basic comprehension: pronoun reference' question types. 'Basic comprehension' may indicate that this item type should target propositional understanding at the clause or sentence levels. However, this item type cited pronoun references in multi-clause sentences, requiring participants to distinguish between multiple noun phrases in preceding sentences. Inferential reasoning at the sentence level occurred most prominently in 'inference' and 'basic comprehension: pronoun reference' question types. 'Basic comprehension' may indicate that this item type should target propositional understanding at the clause or sentence levels. However, this item type cited pronoun references in multi-clause sentences, requiring participants to distinguish between multiple noun phrases in preceding sentences. As a result of this, participants tended to summarise information across sentences by way of explanation. 'Basic comprehension' as understood by the TOEFL test developers encompasses both high and low-level processing. Inferential reasoning appears to be under-specified in IELTS. Positive evidence of this level of processing in IELTS is sparse.

The final section in the findings and discussion chapter considers the main research question which drove the research. This section reflects upon the framework of RE created in the literature review and considers how the findings of the research have been influenced by

the framework and the extent to which the framework has driven methodological innovation in the study.

4.6. Can an evidence-centred framework of reverse engineering be used to develop representative cognitive test specifications for the purposes of test comparability?

This thesis has demonstrated that an evidence-centred approach to RE (Almond et al, 2002; Mislevy et al, 2003 was successful in deriving representative cognitive specifications from participant interactions with the IELTS and TOEFL tests. This section reflects upon the approach to RE cognitive test specifications in the thesis and the implications of the research design in terms of the methodological decisions, conceptual understanding of RE and the wider language testing literature.

The purpose of the thesis was to develop a robust theoretical framework of RE for a wide range of research agendas by diverse stakeholders. The framework is therefore designed to encompass 'straight' RE in order to produce descriptive test specifications so that stakeholders can develop descriptive specifications to produce equivalent test instruments as well as 'critical' RE in order to reflect upon and critique the understanding of the construct embedded in test design. The emergent RE framework was presented in the literature review (figure 2.4), and is reproduced here for quick reference:

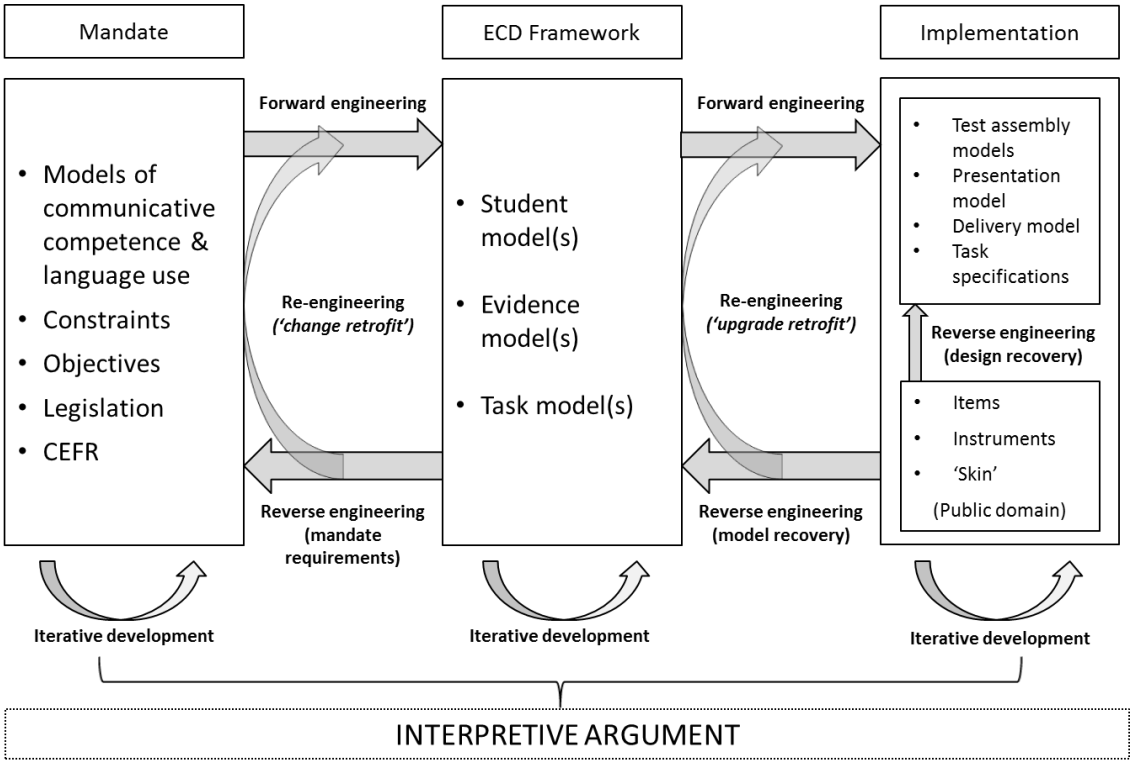


Figure 4.19. An evidence-centred framework of reverse engineering.

The framework of RE integrates the three main layers of architectural documentation (Fulcher and Davidson, 2009: 127), corresponding elements of the CAF (Mislevy et al., 2000; Mislevy et al., 2003; Mislevy and Riconscente, 2005) and the relationship between RE, forward engineering and reengineering. The framework contains three columns. From left to right they represent traditional broad stages of 'forward engineering'. An institution, government agency or other organisation identifies a need to make decisions about individuals related to some aspect of language proficiency and decides that an objective assessment acting in accordance with existing legislation and best linguistic theory undertakes assessment design, construction, field testing and then launches the test into the public domain (mandate, ECD framework and implementation). Each of these stages includes iterative development by the developers to improve the coverage of the target domain or improve statistical qualities or strengthen validation arguments. In conducting RE to make a new specification or to uncover construct-relevant design principles, the engineers will need to consider who the specification is for and why. Different stakeholders will require different information. The goal was to forward a standardised approach to RE to

provide strong epistemological grounds for comparing outcomes for different research instruments subjected to RE.

A formalised approach to RE focusing on the six ECD stages of test development (figure 4.19) benefits item writers and a variety of stakeholders with an intrinsic interest in a test such as IELTS or TOEFL. Test score users who wish to scrutinise a suitable 'product' for their admissions or gatekeeping procedures will want to ensure that successful completion of a test elicits abilities which are sufficient for participation on their academic programmes. RE may also be a valuable method of developing positive washback by providing cognitive information to teachers and learners which may be used in teaching and learning. Teachers and learners will become more aware of their capabilities and the requirements of test items, and the strategic decision-making associated with correct outcomes. Teachers are not always equipped with the skills to help students to develop their monitoring accuracy and metacognitive appraisals (Kostons et al, 2012). Language teachers may not have the means to effectively provide cognitive feedback to their students (on whether their students are using appropriate search strategies, or know the appropriate unit of information (word; clause; sentence; multiple sentences; paragraph; text) to answer specific items. For this reason, a formal framework of RE which encompasses detailed information about requisite strategic decision-making and levels of cognitive processing associated with an item type should be developed for teacher training and classroom use as part of high-stakes test preparation.

This section summarises the main concepts of the ECD approach to RE and considers how each has contributed to the findings, and how this focus led to a more in-depth and comprehensive comparative analysis. Section 4.6.1 specifically considers the importance of ECD to the RE framework. Section 4.6.2 forwards a new conceptual understanding of RE based on the experience of using the framework in the thesis, and how the development of the framework has led to greater understanding of the concept. The thesis has culminated in a reconsideration of RE, of which this thesis represents the most comprehensive study to date. Sections 4.6.3 and 4.6.4 reflect on how the framework influenced methodological decisions and how the study can have important implications for future studies and the

wider language testing literature. An evidence-centred approach in the thesis has resulted in methodological innovations which would otherwise would not have occurred.

4.6.1. Reflections on the utility of ECD-based reverse engineering in the present study and relevance to wider literature

This section provides a reflection on the use of the RE framework to address the question of test comparability, specifically how an ECD approach influenced the research design. This study has demonstrated that evidence-centred design (Mislevy, 2003) offers a clear framework for a research agenda of RE. Secondly, this section situates the study in relation to wider testing literature. RE comes under the umbrella of 'validation studies', as it represents a clear attempt to explicitly identify aspects of the construct embedded in the test. Third, this section provides reflections on the methodology used in the study. The study has also introduced innovative changes to stimulated-recall interviews in order to address the requirements of a robust interpretive argument and the limitations of the methodology outline din the literature review. These changes are of particular interest for future studies which involve SRI, such as eye-tracking.

The introduction noted the established definition of RE as "the creation of specifications from representative sample items" Davidson and Lynch (2002: 41). Authors who have discussed RE have not forwarded a theoretical framework within which RE might operate, how relevant research questions may be composed or a methodological procedure on the basis of these requirements. This study has specifically addressed these gaps in the literature, and demonstrated what a comparative study based on the principles of evidence-centred RE looks like. Researchers conducting RE must pre-specify the architectural layers at which the RE will operate in accordance with the research goals. Reverse engineering different layers of the test architecture will reveal both different data and different amounts of data. RE in the present study operates within Mislevy's (2003) conceptual assessment framework (CAF).

Critical RE uses the publicly available test material to uncover the elements of the ECD framework. Research questions in this study encompassed the evidence, student and task

models. The aim of the student model in RE is to identify what is being measured. In critical RE, the student model is expressed in terms of the cognitive processes encoded in the individual items that are the product of deeper engagement with individual items types. Task models state which cognitive processes and completion strategies are required for individual task types. The interplay between the input material and the questions reveals something about the procedure for procuring the correct answer, and thus, how construct-relevant cognitive processes have been conceptualised by the test developers. The evidence model considers what data is needed in order to make claims about tests and how this data may be analysed to identify construct-relevant attributes. In this study, observable strategies form the locus of attention for the test takers and researchers. Individual moments of engagement reveal instances of cognitive processing that the researcher can claim are associated with that item type, if they contribute to the test taker responding correctly to that item.

The question underlying the evidence-centred approach to RE is 'how can we make RE more objective?' Using ECD highlights the premise that the tests can be characterised as hierarchical structures (student, task and evidence models cited in this thesis). Researchers should be able to start at any level of that system and consider individual levels in the hierarchical order of that system by repeated application of that process. ECD is especially useful for emphasising the evidential connection between data collected and claims subsequently made on the basis of that data, such as that based on Toulmin's (2003) structure of argumentation. Robust methodology needs to be able to be applied to any level in the hierarchy structure in order to be able to uncover internal particulars of the elements that are representative of that level. Applying a RE framework to multiple architectural levels reveals how the levels interconnect. RE conducted at different architectural layers will also produce differing specification documents.

The outcome of critical RE is a cognitive specification matrix, acting as a blueprint which may be sampled by test developers seeking to measure specific cognitive processes (Buck, 2001). The test specification blueprint provides validity evidence for how the cognitive processes relevant to the construct have been operationalised. The specification should provide a complete chain of inferential reasoning between the observed variables that the test

requires for successful completion and the construct definition, which in this study is provided by Toulmin's (2003) interpretive argument. A test specification at this fine level of granularity shows how these processes have been operationalised at the item level, to the extent that a researcher may create items that have congruence with existing items (Fulcher and Davidson, 2007: 377) at the cognitive level, which this study has successfully demonstrated for the reading components of the IELTS and TOEFL iBT.

We cannot say that a specification which includes a carefully-worded definition of the construct is necessarily more 'evolved' than one which does not, as test developers may have decided to publish such information in an alternative document. There is no clear-cut recipe for what a specification should include, with alternative models proposed by different scholars over time. The distinction between straight and critical RE addresses these observations. A general framework of RE encompasses both straight RE, a research project that is not necessarily about getting back to a construct definition and critical RE. Critical RE attempts to uncover the developers' understanding of the construct. An ECD approach to RE is able to encompass both of these research agendas.

The agenda of developing an RE laid the mandated theoretical and methodological item analysis with authentic test taker engagement with both tests at a very finely-grained level. This study has prioritised the cognitive components of the two tests in elaborating a cognitive specification. This cognitive level of engagement does not presuppose that other elements of test or item design are unimportant, rather that the study has purposefully limited itself to those elements deemed sufficient to create cognitive specifications for the tests and item types. This was undertaken to demonstrate that an ECD-operationalized RE framework was capable of producing claims about high-stakes tests beyond descriptive design principles. RE, conducted with different methodological frameworks, is a suitable framework for producing data congruent with the research questions or agenda that drove the RE for a variety of stakeholders working in different contexts. As a result of the disparate aims and methods by which RE may be conducted, this study proposes a new conceptual understanding of RE.

4.6.2. A new conceptual understanding of reverse engineering

The current definition of RE is "the creation of a test specification from representative sample items" (Davidson and Lynch, 2002: 41). Considering that the outcomes of RE will differ depending on the aims of the reverse engineers, the existing definition of RE is insufficient to encompass a broad spectrum of studies that can involve 'test disassembly' and the scrutinising of a test instrument for the purpose of linking item design elements to the construct of interest more directly. Although such studies discussed by Davidson and Lynch (2002) may come under the rubric of RE, the definition is not exhaustive of all RE possibilities. This study therefore proposes a new definition of RE in language testing to reflect these broader possibilities. RE can instead be defined as:

> *The principled, critical analysis of one or more of the architectural layers of a test, for the purpose of uncovering and elaborating the design principles or understanding of the construct which underlies that test.*

This definition encapsulates RE conducted at any architectural layer and opens up new analytic and methodological possibilities which can come under the umbrella RE. It also demonstrates that re-engineering of a test that occurs during test development and conducted by test developers may also be a form of RE. A framework of RE can be beneficial to test developers as well as independent researchers. Test development has long been recognised to be 'recursive' or 'iterative' (Weir, 2005). Data from field testing in the form of classical statistics, item-response modelling and differential item functioning (DIF) is used to evaluate the performance of items and item types in terms of facility, theoretical model fit and discrimination. This data provides information regarding which items to accept or rewrite, but does not provide information regarding the nature of the problems in relation to specific items. A form of RE as proposed in this study would provide additional qualitative data which would be useful for item writers to determine how an item may need to be re-written, guidance on how to write future items, or a research space to ensure that the items are eliciting the required level of cognitive processing in addition to producing statistical data in line with the test development procedures.

4.6.3. Reflections on the methodology used in the study

The study has revealed that RE is subject to the limitations of any exploratory framework, and will also be subject to the same advantages and disadvantages of any methodology adopted for the RE. Verbal behaviour (participant responses) is used as evidence to explain the observable behaviour of the participants in the video recordings. Responses to the test items coupled with participants' verbal explanations of their actions are used to infer the highest level of processing that participants used (the level that was necessary to complete specific items). The approach is rich in observational evidence. The use of video allows for an 'objective' claim regarding participant behaviour. Each verbalisation can be traced to an observable instance of behaviour, and thus provides the explanatory framework for interpretation. The emphasis in the interviews was on orderly, directly observed behavioural patterns.

4.6.3.1. Weaknesses of the approach

The research was carefully designed to mitigate the known drawbacks of SRI. Nonetheless, several drawbacks associated with the approach have affected the emergent data. Timing was not an explicit consideration in the current study, despite the likelihood that it will impact how test takers engage with the test and the corresponding level of processing. Weir (2005: 65) notes that "time constraints for the processing of text and answering the items set on it will affect the nature of what is being tested". Allowing greater time than would be allowed in a live test may have provided participants with the space to form a mental representation of the text which they would not otherwise have been able to do. This would skew the findings towards the higher level processes in Khalifa and Weir's model. However, the counter argument is that high ability test takers, which this research purposefully recruited, would be able to form mental models of the text without the additional time allowance. The emphasis in the study was on obtaining as many correct responses as possible to determine a complete cognitive specification matrix of both tests to demonstrate the value of RE and for direct comparability purposes. If the research was conducted with lower level participants, their inability to form mental representations of

the text would not be assisted by extra time. Adjusting the timing of a test will likely have the biggest impact on middle ability participants who may already feel time pressure and small changes in timing may alter their scores by a wider margin relative to higher ability test takers, who formed the research participants for this study.

In Narens' (1990) model of metacognition (knowledge about and regulation of one's thinking), there are two levels of mental processes: cognitive/object and metacognitive levels. One possible implication of the methodology is the blurring of the distinction between automatised or routine metacognitive strategizing related to unconscious actions on the part of the test taker and the overt, metacognitive strategic decisions. Showing participants a video of their performance may cause them to reflect upon their actions and to elaborate what was previously automatized, giving the impression that this was a conscious strategy on the part of the test taker. Additionally, within the stimulated-recall interviews, there is no way to discern cognitive and metacognitive statements. By definition, statements made about cognitive processes that occurred during the performance displayed in the video have been reconceptualised by the participant when elaborating during the interview. This will result in a preponderance of inaccurate reporting verbs such as 'I decided…' to refer to instantaneous moments of cognitive processing during task completion.

A third weakness which relates to the timing is the order in which participants completed the items. The methodology required the participants to move through the test in sequence. As a result, they may not have completed the items in the order in which they would in a live test. This is discussed in section 4.3.2.1 (in relation to the IELTS multiple-choice items), as participant 1 stated that she would have bypassed the multiple choice items to examine easier items before returning to the multiple-choice items. No other participants mentioned this issue, however. In future studies, participants could be asked if they completed the items in the order in which they would during a live administration to gauge how the method has impacted on the findings.

Fourth, verbalisations were made in response to stimuli presented to the participants. Verbalisations in the study were therefore highly situated. Verbalisations in the present

study relate to instances of observable behaviour in the videos and on participants' answer sheets. In other words, they will be focused largely on reflections made during moments of overt strategic management. Similarly, it is possible to 'overload' the participant (over-stimulate) potentially resulting in confused or erratic verbalisations which cannot be triangulated or compared with other participants to ensure reliability or to sustain an argument in relation to specific items. For this reason, it was necessary to develop a coding algorithm to justify and defend the use of codes from Khalifa and Weir's model to individual verbalisations. The development of this algorithm facilitated the transition of Khalifa and Weir's processing core from a model of reading to a research instrument. This algorithm was applied to all verbalisations that were gathered in the data collection to form the basis of the specification matrix and cognitive specifications for both the IELTS and TOEFL items.

As a result of the specificity of the approach, the study did not employ multiple raters. This is justified primarily on the basis that determining the applicability of the coding scheme developed from Khalifa and Weir's formed part of the findings. Utilising multiple raters would have resulted in an additional layer of interpretation to ensure that raters had applied the coding in a consistent manner. An algorithm was developed in a grounded approach of engaging with participant verbalisations, comparing them to the video evidence and participant responses to test items to mitigate the issue of no measure of inter-rater reliability being included in the study.

The nature of the present study remains exploratory given the novel nature of the application of video-stimulated recall interviews and innovative coding framework as part of an agenda of RE. Findings need replication with other IELTS and TOEFL tests and other nationalities, and different ability participants. Replication of findings in a larger-scale study using the procedures developed in this study with input material that has been through the full editing stage of the test developers to see whether the findings hold true.

4.6.3.2. Advantages of the approach

However, the findings also reveal several distinct advantages of the approach. The advantage of the technique lies in the nature of the stimuli and interview structure allowing participants to formulate and structure their own memory recall. The video acts as a vivid stimulus of short-term memory. Adult L2 learners possess well-developed and stable cognitive capabilities, whereas their strategic management may change rapidly depending on their familiarity with task types and instructions. Evidence from participant verbalisations and resulting claims about the two tests suggests that the verbalisations are good sources of data to make claims about cognitive processing at specific moments in the test. Differences in performance ability by participants are more likely to be due to different levels of linguistic competence rather than strategic competence (Field, 2013). Nonetheless, the participants are clearly aware of their own information processing, even in an artificial and unfamiliar (Phakiti, 2007) SRI context. Verbalisations made compelling data on which to make claims about the nature of reading each test 'encoded' in its design.

Verbalisations are therefore superior to strategy questionnaires to elicit mental processes. Test takers are generally aware of the types of strategies that may be deployed to assist them in responding to different items, and are able to report accurately that they have attempted to use a specific strategy. However, asking participants to report the cognitive strategies they use in a test is likely to produce volatile data that will only have a low correlation with an instrument that asks participants what cognitive resources they typically deploy. Phakiti (2008) argues that cognitive strategy use in reading was less stable over time than metacognitive strategy use. How participants perceive they use their cognitive resources is not necessarily a good predictor of what resources they actually deploy for any given item or test. Asking participants to report the cognitive strategies they use in a test is therefore likely to produce volatile data that will only have a low correlation with an instrument that asks participants what cognitive resources they typically deploy. This is an advantage of the current technique over a questionnaire instrument asking participants to self-report their processing. Findings are based on interpretations of participant actions, coupled with their description of how they engaged with specific portions of the text at given moments. Online metacognitive strategies (which test takers are aware of and are able to deploy) work hand in hand with online cognitive strategies to accomplish the test tasks.

Participants may go to a text and commence a search strategy, but they may be holding information relating to more than one item in their working memory. If they are presented with strategy questionnaires pertaining to individual items, they may only respond that they use a strategy with a single item, which would be an inaccurate representation of their mental processing, or they may report the strategy for both items, which may result in that strategy being over-reported relative to others which were used at the same time but in relation to only one item. Counting instances of strategy occurrence may present a distorted impression of the importance of some strategies relative to others, which is in actuality an outcome of method effect. Counting instances of strategy use tells us nothing about the quality of the strategies used. There is no general consensus in the literature for analysing verbalisations made on the basis of visual data in video stimulated recall interviews. The methodology offered in this thesis is offered as way forward for subsequent studies based on similar methodological approaches. The requirements of the RE framework demanded a strong inferential argument regarding linking claims to specific verbalisations. The argument-based approach provided the basis for the inferential reasoning.

The RE framework was the driver of methodological innovation in the study. The evidence model prompted a rich and detailed examination of how verbalisations reflected levels of engagement. This methodology is an example of researcher-mediated data, insofar as the instances of participant behaviour included in the timeline have been identified by the researcher in the context of replaying the video. There was therefore a heavy focus on recorded observable elements. This served two functions – first, it provided 'hooks' for the researcher to latch on to stimulate the participant's memory. These could be used to prompt the participant to verbalise why they had undertaken specific actions. Moments were directly observable so required no specific explanation to the participants to enable them to contribute to the interview. Second, the video could be replayed at a later date, allowing the researcher to replay crucial moments of behaviour with subsequent participant verbalisations to aid coding.

Moments in the video in which the participant did not appear to be taking any action indicated that the participant was most likely engaging with some portion of the text in the

form of careful reading. These moments were also cited by the researcher for the participant to verbalise what they were thinking during those moments. The lack of concrete action makes these moments more difficult for the participant to explain. This process was assisted by actions taken by the participants after this period of inaction. These can be linked to the moment of inaction as the participant can be asked why they undertook that action and invited to elaborate. This may prompt them to recall some detail of the text they had just been engaging with, providing greater verbal evidence to identify the level of engagement with the text.

The final part of this section situates the thesis within the wider language testing literature. The thesis has sought to identify key aspects of the construct of reading in English for academic purposes. Therefore, the study can be described as part of a validation argument. The next section considers how RE can be considered as part of an ongoing validation agenda and the link between the methodology chosen in this study and methodology in ongoing validation studies, specifically eye-tracking.

4.6.4. Reflections on the relationship between RE and wider language testing literature

RE, like any other attempt to link a test to its construct of interest, comes under the rubric of validation studies. This study has been predicated on ECD, a framework for linking claims specifically to data via Toulmin's (2003) model of argumentation. Kane (2006) attempts to streamline validation by developing this argument-based methodology to support ongoing validation practices, alongside guidelines provided by the Standards (AERA, 2014), validation studies remain difficult to conceptualise and execute. RE, as conceived in the present study, is subject to the same challenges associated with an evidence-based validation project: "those who are actually responsible for validation almost always require detailed and concrete practical guidance for conducting validation activities" (Brennan, 1998: 7).

The study focuses on just one skill because each language skill is unique and complex (van Patten, 1994) and should be specifically and comprehensively researched (Schmidt, 1995).

Field (2013) notes the important role of strategic competence in L2 listening proficiency because it helps L2 listeners make sense of listening in a real world setting; proponents of metacognitive strategic processing in reading would cite the importance of members of an academic discourse community needing to make decisions about what they read and how they engage with individual texts in order to successfully complete a task, and that reading task in tests are designed to elicit test taking behaviour that is comparable to real-world decision-making processes. "For example, metacognitive learners will engage in self-monitoring and evaluation during their learning, so that they can check whether they have met their objective or satisfactorily completed the assigned tasks" (Cohen 2011).

The literature review critiqued the literature surrounding cognition in language testing for operating without agreed-upon definitions for key terms used by practitioners in the field. Thus, there is some uncertainty when reading and comparing studies that different authors have similar conceptions of the terms that they use. The use of Khalifa and Weir's model provided not only an analytical tool, but also definitions of key terms to work from. Strategic competence refers to a set of metacognitive strategies which regulate the cognitive *core* in Khalifa and Weir's model (cognitive processing). This core represents the test taker's ability to engage with material presented to him or her with the linguistic resources available (language *use*). Too often the conscious strategic management of a task by a test taker is conflated with the cognitive processing of the test taker as they engage with the task. A test task is a simplification of reading activities from the construct. This is accepted in testing because the argument is that this task requires some aspect of the processing competence that is required in the construct, despite the task itself not representing any task the test taker will be required to perform. For this reason, it does not logically follow that the strategic management of that task is necessarily construct-relevant. Strategic competence refers to a number of non-linguistic factors affecting language test performance.

Strategic competence has been recognised as a significant cognitive factor that distinguishes successful test takers from less successful ones (e.g., Bachman & Palmer 2010; Cohen 2011; Phakiti 2003; Purpura 1999; Zhang and Zhang, 2013), but this also does not represent evidence that strategic competence is construct relevant. If strategic competence is a distinguishing factor between successful and unsuccessful test takers, then strategic

competence could represent a form of construct irrelevant variance if it produces 'false negatives'; participants who possess the necessary linguistic resources to complete a task, but not the strategic knowledge to deploy their resources appropriately in that context.

The extent to which *strategic* competence in a test setting matches the strategic competencies associated with successful language learning is not necessarily construct-relevant for a test of academic English, the purpose of which is to make a decision regarding admission to higher education by predict the success of a participant in completing their academic programme on the strength of their English proficiency, not to make a statement about their future success in language learning. Test taking strategies that are analogous to language learning strategies will be of great interest in a diagnostic setting, in which the purpose of the test is to identify weaknesses in both the test takers language proficiency and their metacognitive decision making to determine how they can improve and manage their own learning.

4.6.4.1. Utility of the methodology for eye-tracking studies exploring cognitive validity

Recently in language testing, there has been a shift towards using eye-tracking as a methodology for identifying cognitive processing in relation to specific item types and to incorporate this data into validation arguments (Bax, 2013; Brunfaut and McCray, 2015; McCray, 2014). Eye tracking is the measurement of eye activity in relation to an on-screen stimulus. Software can measure fixation (duration of gaze at a single point), saccades (forward eye movement) in pixels and regressions (backward movements). These metrics provide information about participants' engagement with source material. Patterns can be compared between successful and unsuccessful participants to identify what necessary metacognitive and cognitive processes are required to complete an item. Bax (2013) argues that eye-tracking is useful for identifying metacognitive strategies by participants and for differentiating between low level processes (lexical access and syntactic parsing). Evidence for the efficacy of eye tracking to reveal and distinguish between higher level processes is implied by the absence (or limited duration) of fixations on specific lexical items (Bax, 2013; McCray, 2015).

Eye-tracking is usually accompanied by stimulated-recall interviews to provide qualitative data to reinforce the interpretations of the eye-tracking data. The methodology developed for the interviews in this study is of great benefit to eye-tracking studies. In eye-tracking, participants are often presented with gaze tracking videos, in which a small dot marks the location of gaze, with a small path indicating movement. Any research of this sort is limited by the fact that performance appraisal in the form of a reflection on what cognitive actions participants undertook at specific moments are treated as *conscious* activity, while in fact some may operate at an unconscious level as *automatized* movements. Fixations are often measured in milliseconds and as such are difficult for participants to assign meaning to. Similarly, highly routinized or practised metacognitive strategies such as planning and reviewing (e.g. eye tracking metrics associated with recasting, or moving back in the text) can become automatised at the *object level*, and are part of a test taker's executive test management at the *macro* level, suggesting that metacognitive strategies can be explicit at the meta level, but implicit at the object level. This suggests that participants shown their eye movements on a screen are being asked to recall a procedure that is heavily automatized and thus will not be able to produce a meaningful verbalisation. The level of conscious and non-conscious awareness of metacognition is not easy to resolve. Efklides (2008) pointed out that the association of metacognition with consciousness is no doubt absolutely necessary to help researchers understand how people take control of their cognitive activities, especially when automaticity fails.

Inferences linking eye-movement and fixations to specific processing are dependent on the 'eye-mind hypothesis'. This is the assumption that individuals are thinking about whatever they are looking at during specific moments of fixation has been criticised in the psychological literature (Anderson et al, 2004). It is possible that participants are examining one section of text while reconsidering another in their minds. Ultimately, the only way to access this information is through verbal interaction with the participant – eye tracking remains dependent upon inferential reasoning of what the participant is thinking. Video stimulus provides concrete moments of participant action which are relatable. Lower level processing can be observed by the eye tracking as these are specifically linguistic skills. Above sentence processing refers to higher order cognitive organisation of ideas emanating

from a text. Executive test taking skills are organisational rather than linguistic abilities for managing a task and for relating information across sentences (forming coherent meaning of the text). These processes are more difficult to observe. The methodology adopted in this study is one means of accessing these higher-level processes.

This concludes the findings and methodology chapter. The final chapter of the thesis is the conclusion. In this chapter, the main findings and contribution to knowledge will be outlined. The conclusion brings together the disparate parts of the thesis and summaries the research journey that this thesis represents, from conception to overall outcomes.

Chapter 5. Conclusion

5.1. Introduction

Universities in English-speaking countries require evidence of English language proficiency prior to enrolment, which is presented in the form of a test score. This is obtained by taking a test provided by one of the large test developers such as Cambridge Assessment (who produce the IELTS test) and ETS (who produce the TOEFL iBT). If participants present evidence from different test developers, university admissions staff need to be sure that test scores can be compared directly. Comparability studies to date have focused on comparing scores obtained from test takers who have completed multiple English language tests. The thesis was inspired by the need to provide strong empirical evidence that the tests are able to be compared, which must occur before consideration of the technical aspects of score comparability. Therefore, the focus of the thesis was on uncovering working definitions of the construct underpinning the design of the IELTS and TOEFL tests. A comprehensive analysis of both tests was beyond the scope of the thesis. Therefore, the focus of inquiry was the reading sections. Despite the development and extensive changes that have been made to both tests in the past two decades, Moore et al (2007) note that both tests have retained an independent reading section. Reading is an essential competence in higher education, warranting further investigation.

Different tests may claim to measure the same construct, but will be designed and delivered in very different ways. There is little publicly-available information from test developers regarding the working definitions of the construct of 'reading for academic purposes'. Information regarding how a construct has been operationalised by test developers is often found in a test specification. Test specifications are blueprint documents (Fulcher and Davidson, 2007) which are used for iterative test development. They are proprietary documents which are not available for public consumption. Gaining access to specifications would provide stakeholders with more information in order to make decisions about which test is suitable for their purposes if they have a different conception of the construct of interest. The aim of the thesis was therefore to uncover the conception of 'reading for

academic purposes' which underpins task design in the two tests and to compare the extent to which they have a shared conception of the construct.

RE offered a solution to the problem. RE refers to the procedure of creating a test specification where one is unavailable. This is an idea that that has been included in some influential works (Davidson and Lynch, 2002; Fulcher and Davidson, 2007) in language testing, yet had received scant attention beyond general discussion. A secondary impetus for this thesis was therefore to develop a systematic approach to test analysis which would produce a representative construct-oriented test specification.

This chapter concludes the thesis. It provides a summary of the whole thesis based on the twin aims here of comparing the reading components of IELTS and TOEFL tests in terms of the respective developers' conception of reading for academic purposes, and developing a systematic approach to the concept of RE. This chapter mirrors the progression of the thesis, so begin with an overview of the development of a systematic approach to RE, which acts as the analytic framework, and the contribution that this has made to the language testing literature. The second section considers how the new analytic framework of RE led to the development of the key research questions and directly influenced decisions regarding an appropriate methodology, and led to methodological innovations which were made in order to increase the robustness of the findings. The third section presents a summary of the findings relating to each research question. The final section considers the contribution to the literature that the thesis has made in two ways; test comparability research and the construct of L2 reading.

5.2. Theoretical contribution to the development of reverse engineering

The study succeeded in developing a theoretical framework of RE using two principal sources. The framework was developed by reviewing and incorporating literature from systems engineering and a procedure of test development known as 'evidence-centred design' (ECD). RE is not a new idea in systems engineering and contains striking analogies to language test development, which this thesis utilised.

This thesis has delineated between different 'types' of RE based on specific research agendas. Systems engineering identifies 'cloning', which implies the duplication of an existing system, and 'surrogacy', "gain[ing] a sufficient design-level understanding to aid maintenance, strengthen enhancement, or support replacement" (Chikofsky and Cross, 1990: 14). This served as the inspiration to add to the literature on RE by clarifying the difference between 'straight' and 'critical' RE, which was previously under-developed. Straight RE is the process of analysing test items to determine design principles, such that equivalent items may be created to the same specification. Critical RE goes beyond inference of guiding language, is the process of analysing items to determine whether they measure an aspect (or aspects) of the construct which we require them to. This study was concerned with the latter form of critical RE, which can provide crucial validity evidence that the tests measure what they claim to.

This study has also provided an example of the theory and practice of RE as distinct from *retrofitting*, a procedure which has already received theoretical attention in the literature (Fulcher and Davidson, 2009). Chikofsky and Cross (1990) also helpfully distinguished between RE and retrofitting. In systems engineering, the latter is explicitly conducted by the original developers in relation to explicitly stated requirements and research goals. *Upgrade retrofitting* realigns the test instrument with its originally-stated purpose. *Change retrofitting* refers to modifying an instrument to meet the needs of a different test purpose (Rekoff, 1985: 124). The new framework (see literature review, figure 2.4) adopts this distinction, and allowed for the independent theoretical development of RE as a related but distinct concept from retrofitting.

The content of the new RE framework was informed by evidence-centred design (ECD). ECD, with its modular approach to test development and argument-based approach, fitted the RE agenda perfectly. A formalised approach to RE utilised the conceptual assessment framework (CAF) of ECD stages of test development, including the student, evidence and task models. The new framework of RE integrates these three main layers of architectural documentation (Fulcher and Davidson, 2009: 127), additional elements of ECD relating to test design, the presentation, delivery and test assembly models (Mislevy et al, 2003) which do not feature in the current thesis as they relate more to 'straight RE' than 'critical RE'

(Fulcher and Davidson, 2009), and is underpinned by an interpretive argument (Toulmin, 2003). The framework also depicts the relationship between RE, forward engineering and reengineering within the framework of RE offered by Chikofsky and Cross (1990).

The framework is a useful tool for evaluating the evidence for a validity argument and creating a variety of specification types, from item type descriptive guidelines through to a grand table of cognitive specifications. This thesis has demonstrated that the framework can be successfully utilised to address research questions regarding cognitive information about the construct contained within two tests of English as a foreign language. The framework contains three columns. From left to right they represent traditional broad stages of 'forward engineering'. An institution, government agency or other organisation identifies a need to make decisions about individuals related to some aspect of language proficiency and decides that an objective assessment acting in accordance with existing legislation and best linguistic theory undertakes assessment design, construction, field testing and then launches the test into the public domain (mandate, ECD framework and implementation). Each of these stages includes iterative development by the developers to improve the coverage of the construct or improve statistical qualities or strengthen validation arguments. This general framework of RE encompasses both straight RE (the right-hand column); a research project that is not necessarily about getting back to a construct definition, and critical RE, which attempts to uncover the developers' conception of the construct of interest (the central column).

This research agenda has fulfilled the requirements of the definition of 'critical RE' as forwarded by Davidson and Lynch (2002) and Fulcher and Davidson (2007) by illustrating the process in the lower right-hand arrow ('reverse engineering: model recovery') (figure 2.4) to obtain sufficient information about the student, evidence and task models to produce a representative cognitive specification of the reading sections of IELTS and TOEFL.

The aim of the student model (SM) in RE is to identify what is being measured. In straight RE, engagement with the instrument is sufficient to uncover the design principles associated with particular items. In critical RE, researchers need to engage with both the instrument

and the test taking population to identify which mental processes are encoded into item design. In critical RE, the student model is expressed in terms of the cognitive processes encoded in the individual items that are the product of deeper engagement with individual items types. The task model (TM) focuses investigation on each task type, to understand how that task contributes to the overall student model. This addresses the question of which cognitive processes are required for completion of that item type. For receptive skills (such as reading) the interplay between the input material and the questions reveals something about the procedure for procuring the correct answer, and thus, how the skills identified in the construct have been conceptualised by the test writers. The evidence model (EM) explicitly considers what data is needed in order to make claims about individual test instruments and how this data may be analysed to identify construct-relevant cognitive processes.

The RE framework is now established as the basis for further investigation, either with RE as part of an overall validation argument, or further test comparability in other skill areas or other test instruments. The RE framework provided the impetus for focusing on aspects of test comparability that otherwise would have remained under-investigated. Specifically, the evidence model required an extensive focus on methodology in order to link data to claims as part of an interpretive argument. This is discussed in section 5.5, which summarises the methodological innovations driven by the use of the RE framework. The next section summarises the research questions which emerged from the framework and the findings in relation to each of them.

5.3. Overview of methodology used in the study and emergent research questions

Two complete reading tests were selected for the study, one IELTS and one TOEFL. Each is composed of three texts with associated questions. The IELTS test contains forty items and the TOEFL test contains thirty-eight items. These were selected on the basis that they were representative of the reading tests commonly produced by each of the test developers. Texts from eight complete IELTS and eight complete TOEFL tests (twenty-four texts from

each) were submitted to Coh-Metrix[7] and VocabProfile (Cobb, 2013)[8]. These are text analytic software programmes which produce a range of readability statistics. Fourteen metrics recorded statistically significant differences between IELTS and TOEFL. For each of these metrics, means and standard deviations were calculated across the texts for each of the tests. One test was chosen from IELTS and TOEFL tests analysed. The test selected would be that which recorded the majority of the metrics within one standard deviation of the mean, thus ensuring that this texts were as representative as possible.

Each of the tests were divided into three parts, consisting each of a single text and associated items, resulting in six testlets. Six participants were recruited for the study. Each participant completed two testlets, one IELTS and one TOEFL. Participants 1-3 completed the IELTS testlet in their first session, with the TOEFL testlet in their second session. Participants 4-6 complete the same testlets, but in the reverse order to account for method effect. As they completed the test, participants were video-recorded, with the camera aimed at their hands and test papers. When the participant was satisfied they had finished each segment, they informed the researcher, who was seated next to the participant. The camera was connected to a laptop computer and the video of the performance immediately uploaded. Participants were then asked about what they were thinking at specific moments of the video. The video, participant responses, annotated questions and text formed stimuli for participants to talk about their item-completion processes. This is known as a stimulated-recall interview (SRI). Video was used to address known methodological issues with this type of interview (Gass and Mackey, 2000), which are discussed in section 5.5. Observable interactions in the video were labelled 'strategies' as they represent the conscious decision-making procedures of the participants. Research question 1 specifically asks what decision-making procedures can be identified and whether there are any similarities and differences between how participants engage with IELTS and TOEFL. This is the 'evidence model' of the RE processing core.

An analytical framework based on the cognitive processing core of Khalifa and Weir's reading model (2009) was applied to participant verbalisations. This core represents a

[7] http://www.cohmetrix.com/
[8] http://www.lextutor.ca/vp/comp/

hierarchical progression of engagement with text. Lower levels represent local engagement. As one moves up the core, the cognitive load increases. These levels were transformed into a coding framework, with the hypothesis that each of the levels would be able to be applied to participant verbalisations in SRI if participants received sufficient stimuli. Each of the levels was given a code and verbalisations were analysed for evidence of the highest level of processing that could be inferred on the basis of participants' explanation of their actions and responses to specific items. During the analysis of the data, several of the categories were divided as it became apparent that the data was sufficiently finely-grained to allow for a greater number of codes. Establishing propositional meaning and inferencing were divided into two further sections:

	Level of processing	**Code**
Higher-level processes	**Creating an intertextual representation**: construct an organised representation across texts	**[P8]**
	Creating a text-level representation: construct an organised representation of a single text	**[P7]**
	Building a mental model: Integrating new information; enriching the proposition	**[P6]**
	Inferencing: At word/sentence/clause level	**[P5s/c]**
	Inferencing: At word level	**[P5w]**
Lower-level processes	**Establishing propositional meaning**: At sentence level	**[P4s]**
	Establishing propositional meaning: At clause level	**[P4c]**
	Syntactic parsing	**[P3]**
	Lexical Access	**[P2]**

Table 5.1. Final coding scheme for processing core of Khalifa & Weir's (2009) model of reading

The bottom four levels refer to lower-level, or local processing. The upper levels represent higher-level processing; that is, processing of information in terms of ideas, rather than grammatical parsing. Changes to the model represent an addition to our understanding of how to conceptualise cognitive processing in L2 reading, which are summarised in section 5.6. The coding scheme was applied to participant verbalisations using a decision-making

algorithm. Applying the coding scheme to all verbalisations addressed research question 2, which identifies the cognitive processes that test takers use to complete the IELTS and TOEFL tests. Research question 3 highlights any areas of similarity and difference between the IELTS and TOEFL tests. Research questions 2 and 3 relate to the 'student model' of the central core of the RE model. Finally, research question 4 applies a scale weighting to the findings to determine the relative importance of each of the strategies and cognitive processes to each of the item types. Thus, a 'cognitive specification matrix' is produced, which represents how each of the cognitive processes in the tests are embedded in specific item types. This represents the 'task models' of the central RE column. The next section presents summaries for each of the emergent research questions.

5.4. Summary of findings relating to each of the research questions

The following section presents summaries of the main findings relating to each of the research questions in the study. There were four research questions relating to each of the models in the central column of the RE framework (the student, task and evidence models). The research questions were aimed at uncovering and describing similarities or differences between the IELTS and TOEFL tests. Research question 1 related to the evidence model:

What observable test-taking strategies do test takers use when completing IELTS and TOEFL reading tests? Are there any differences in how participants respond to IELTS and TOEFL tests?

Following Bachman and Palmer (1990, 1996, 2010) and Phakiti (2008), this study separated strategic competence from language ability (cognitive processing) for analytical purposes. Participants have command of the test management decisions they make and these may be observed or reported by the participants and researchers. A video-based approach to strategy identification resulted in very finely-grained analysis, allowing differences and similarities between IELTS and TOEFL to emerge.

The evidence model documented the observable interactions made by the test takers when completing the IELTS and TOEFL tests. Through the use of video records of participant

interactions with the reading test and subsequent verbal reports from stimulated-recall interviews, thirty-four individual strategic actions were identified. Twenty-seven of these were directly observable. Seven of the codes required interpretation on the part of the researcher in order to positively identify that they reflected the observed action. Verbalisations with regard to strategic coding were used to confirm that observable behaviour was interpreted correctly by the researcher for these seven codes. For example, frequent instances of a participant moving between parts of the text or the text and item stem would be coded as comparing parts of the stem to the text. These were then used as the basis of subsequent interviews. The video approach to the stimulated-recall methodology resulted in the identification of 1276 individual moments of participant decision-making for the six participants.

Participants recorded a higher frequency of observable strategic actions when taking the TOEFL test than the IELTS test (57 per cent of all observed strategies related to TOEFL). In both tests, participants underlined key noun phrases as focal points for further careful reading. However, in IELTS, participants performed this more frequently in item stems than in the text itself, indicative of greater complexity in the items, as there is more information to process. The most frequently-observed behaviour by the participants in all sessions was underlining (or highlighting by other means) key nouns in the text and question stems. The identification of key nouns was used as a search strategy by all participants as a means of locating the relevant portion of the text. In relation to the TOEFL test, participants then tended to identify and mark further noun phrases that relate to the question stem or options. Participants underlined a much greater proportion of noun phrases in the TOEFL test compared to IELTS.

IELTS test takers moved between the text and the questions more frequently than in the TOEFL. TOEFL items follow a more standardised format than IELTS items and all are presented to the participant in the same, four-option multiple choice format, prompting fewer revisits to the question stem. For IELTS, items are grouped per item type, with one set of instructions for each. ‘Matching heading’ types share one stem, and have short options. For this item type, participants tended to search texts for relevant phrases. Different item types can clearly influence how participants engage with the text and therefore their level of

reading. Further evidence of this is the greater frequency of eliminating options in TOEFL than IELTS. This is also likely a form of method effect as the greater number of multiple-choice items would lead to an elimination strategy.

The TOEFL test recorded more instances of observable careful reading with the text than IELTS. Fifteen codes from the coding scheme align with the definition of 'careful reading' from Khalifa and Weir's (2009) reading model. Twelve codes align with 'expeditious reading', representing 1085 of the 1276 individual codes. Of the total number of careful reading codes, 425 (61.6%) are linked to TOEFL rather than IELTS, suggesting that participants spent proportionally more time reading the texts carefully in TOEFL than they did with IELTS. The total number of codes identified for IELTS was split almost evenly between careful and expeditious reading. Of the total number of expeditious reading codes (IELTS and TOEFL), 260 out of 395 (65.8%) relate to IELTS. This is in contrast to the findings of Weir et al (2009a), who also studied strategy use by IELTS test takers. Their participants self-reported more than three times as many instances of reading carefully as reading expeditiously, suggesting that the methodology chosen has a significant influence on the outcomes of strategy-based research. The robustness of strategy-based research is discussed in section 5.5.

Research question 2: Which cognitive reading processes do test takers use when they complete IELTS and TOEFL iBT reading sections?

Khalifa and Weir's (2009) model of reading was applied to participant verbalisations. Each of the levels was given a code and verbalisations were analysed for evidence of the highest level of processing that could be inferred (using a coding algorithm) on the basis of participants' explanation of their actions and responses to specific items. A level of processing from table 5.1 was applied to each of the verbalisations. The thesis demonstrated that it is possible to infer the level of cognitive processing used by participants based on whether the unit of reference is a single word, a phrase, a complete sentence, an understanding derived from adjacent sentences, an impression formed from engaging with an extended text, or in response to a level of understanding that has been established from multiple points of reference across multiple paragraphs culminating in a

mental model. The level of understanding can be judged by the extent to which the participant rephrases parts of the text in their own words.

Evidence for each of the levels of processing was found for both the IELTS and TOEFL tests, with the exception of the highest level of processing (P8; 'creating an intertextual representation'). This was not unexpected, given that neither test requires participants to answer questions which refer to more than one text. Nonetheless, this suggests that both tests omit a fundamental part of the construct of reading for academic purposes. Analysis of the development of IELTS revealed that the trend has been towards less-integrated sections and the removal of discipline-specific texts. It is therefore extremely unlikely that IELTS will move towards testing inter-textual representation in the near future. In contrast, the revised TOEFL test included 'integrated-skills items', although no item types specifically target the integration of information from two or more texts. As the scope of this PhD is limited to a single skill (reading), the integrated-skills items were not considered for analysis.

Research question 3: Do these processes reveal differences in IELTS and TOEFL iBT test developers' understanding of the construct of interest?

Table 5.2 displays the identified cognitive processes in the participant verbalisations for both IELTS and TOEFL:

Higher/ lower	Cognitive Processing	IELTS	%	Higher/ lower	TOEFL iBT	%	Higher/ lower
Lower	P2	**93**	39.74		**108**	34.84	
	P3	**28**	11.97		**31**	10.00	
	P4c	**40**	17.09		**66**	21.29	
	P4s	**45**	19.23	88.03	**51**	16.45	82.58
Higher	P5w	**2**	0.85		**7**	2.26	
	P5s/c	**4**	1.71		**10**	3.23	
	P6	**20**	8.55		**33**	10.65	
	P7	**2**	0.85		**4**	1.29	
	P8	**0**	0	11.97	**0**	0	17.42
	Total	**234**	**100**	**100**	**310**	**100**	**100**

Table 5.2. Identified cognitive processes for IELTS and TOEFL

As confirmed in relation to research question 2, eight of the nine levels in the coding schema have been identified for both IELTS and TOEFL. The majority of identified cognitive processes were lower level processes for both tests. The TOEFL test recorded a greater number of instances of each level of processing than the IELTS test. It is clear that the pattern of recorded cognitive processes across the two tests is consistent, which suggests that the codes have been consistently applied across both tests, and that they have a very similar conception of the construct of reading for academic purposes. Finally, high level processing in both tests lies primarily in 'forming a mental model' (P6) rather than inferencing or creating a text-level representation.

Some of the higher-level processes in Khalifa and Weir's model are under-represented in both IELTS and TOEFL, including inferential reasoning and forming a text-level representation. This study has expanded upon the definition of inferential reasoning in Khalifa and Weir's model to incorporate both word and sentence-level inferential reasoning, which is discussed in section 5.6. Both types of inferential reasoning identified in this study are identified in both IELTS and TOEFL, although there were more explicit instances of these cognitive processes in TOEFL than IELTS. There was no previous evidence in the literature of sentence-level inferential reasoning for either test, a gap in the literature which this study addresses. Overall, a greater number of cognitive processing coded were recorded for TOEFL than IELTS, in line with the greater number of observable moments of engagement with the TOEFL test than the IELTS test by the participants. The findings in relation to research question 3 represent a significant addition to our understanding the constructs of reading for academic purposes embedded in IELTS and TOEFL.

Research question 4: Are cognitive processes associated with specific item types? Do individual item types target specific processes or do they elicit a range of processes?

This exploration of this research question provided evidence of cognitive processing associated with individual task types. This is part of the task model in the RE framework. Participant verbalisations were analysed in relation to each task type. This research question presented direct analysis of verbal evidence for the cognitive processing in each item type for all task types in TOEFL and IELTS. Each item type from both IELTS and TOEFL was

addressed in an independent section. These findings were then brought together in the form of a cognitive specification matrix (Buck, 2001) in order to directly compare across item types for each of the tests. Weighting was applied to the findings based on the number of items per item type to determine the relative importance of specific cognitive processes for each of the item types. This specification matrix served as the basis for making strong comparative claims between IELTS and TOEFL, specifically in the distribution of higher and lower-level processing across the item types in both tests.

For TOEFL, higher level processes are associated with items that concentrate in the second half of the test. Four basic comprehension item types contain significantly fewer higher level processes than the subsequent four item types (inferencing and reading-to-learn). For TOEFL, the majority of items participants encounter at the beginning of the text will target local reading relating to specific details in the text. TOEFL reading-to-learn items target the top end of the Khalifa and Weir (2009) model and require participants to form mental models of parts of the text. This item type provided strong evidence of these levels of the model, in contrast to IELTS, where evidence of participants' text-level comprehension was less clearly associated with specific item types. For this reason, participants in the IELTS test may encounter an item type that requires processing at a high level as an opening question. IELTS item types each target a range of cognitive processes. Evidence presented here provides justification for IELTS test developers designing tests which do not contain the full range of possible item types in each version. Individual texts may be matched with as few as two item types, and it is usual for a complete test (composed of three texts) to not contain the full range of item types used in IELTS. Inferential reasoning at the sentence level in TOEFL occurred most prominently in 'inference' and 'basic comprehension: pronoun reference' question types. Inferential reasoning appears to be under-specified in IELTS. Positive evidence of this level of processing in IELTS is sparse.

The remainder of the conclusion will summarise the contribution of the thesis to the language testing literature. The thesis has made three main contributions; to literature on test comparability, the construct of L2 reading, specifically the reading model proposed by Khalifa and Weir (2009) and innovations in stimulated-recall methodology. Each will be

summarised in an individual section, before final considerations of the practical implications of the research for stakeholders.

5.5. Contribution of the thesis to test comparability research

Previous investigations into test comparability have focused primarily on the technical approaches to score comparability (ETS, 2010; Holland and Dorans, 2006; Pearson, 2009). Only one major study in language testing has attempted a more holistic comparison of test content (Bachman et al, 1995). This study has not been built on in the language testing literature, despite the choice of tests available to prospective test takers being wider than ever. This thesis offers a proposed approach to test comparability that is principled, robust and encompasses a wide range of analytic possibilities. This study forwards the key message that in order to compare tests, stakeholders must have a clear understanding of how the test developers conceive of and present the *construct* to test takers. If stakeholders and developers do not know what the construct is, then tests cannot be compared sufficiently to be able to make meaningful claims about score comparability.

For many stakeholders, information about the construct embedded in tests is not forthcoming. Information about how the construct has been realised in tests may be regarded as proprietary or commercially sensitive. Limited information may be available on publicly-facing websites about required skills, although may be insufficient to make decisions about the suitability of tests for specific purposes by individual test takers or educational institutions. Information regarding the realisation of a construct may be found in test specification documents containing data on cognitive processes which are relevant to item and test completion. Therefore, the thesis was predicated on devising and implementing a procedure of RE for test comparability purposes. RE does not explicitly have a comparability agenda, although it is one legitimate use of RE. To support multiple research agendas, a new definition of RE is offered to reflect the range of analytic and methodological possibilities which RE can encompass:

The principled, critical analysis of one or more of the architectural layers of a test, for the purpose of uncovering and elaborating the design principles or understanding of the construct which underlies that test.

This definition was formulated as it encompasses the analysis conducted within this thesis, while also laying the groundwork for future studies to be conducted within a framework of RE. The thesis has identified the following purposes for which RE may be used:

1) The principal purpose of RE is to codify a series of skills and procedures for rapidly gaining access to information about the underlying principles of a test without the benefit of guiding documentation (specifications).
2) Identifying the construct of interest in relation to different tests and item types, directly linking the construct to the design and make-up of that test.
3) RE provides evidence that specific item types are capable of eliciting specific cognitive processes and therefore those processes that the developers consider important to the construct of interest.
4) Elaborating the cognitive distinctions between items, between successful and unsuccessful responses and therefore the required level of processing to produce a successful response.

RE as informed by evidence-centred design requires robust design principles. Using ECD highlights the premise that the tests can be characterised as hierarchical structures (student, task and evidence models cited in this thesis). Researchers should be able to start at any level of that system and consider individual levels in the hierarchical order of that system by repeated application of that process. ECD is especially useful for emphasising the evidential connection between data collected and claims subsequently made on the basis of that data (such as that based on Toulmin's structure of argumentation). Robust methodology needs to be able to be applied to any level in the hierarchy structure in order to be able to uncover internal particulars of the elements that are representative of that level. Applying a framework of RE to multiple architectural levels reveals how the levels interconnect. RE conducted at different architectural layers will also produce differing specification documents.

5.5.1. Practical implications of the research

The purpose of RE is to produce a detailed specification for the consumption of stakeholders such as test takers or educational institutions. Stakeholders can then make decisions regarding which of the tests are most suited to their purposes. This was achieved in relation to both the IELTS test and the TOEFL test. A formalised approach to RE focusing on the ECD stages of test development (student, task and evidence models) benefits item writers and a variety of stakeholders with an intrinsic interest in an instrument such as test who wish to scrutinise a suitable 'product' for their admissions or gatekeeping procedures as well as independent researchers with an academic interest in the test under scrutiny.

The outcomes of the RE are presented in a cognitive specification matrix inspired by Buck (2001), who argues that a specification matrix acts as a 'cognitive rubric' so that individual items types may be sampled and fed into an instrument to produce a test that explicitly targets aspects of the construct. The specification matrix may foster positive washback by providing cognitive information to teachers and learners which may be used in teaching and learning. Teachers and learners will become more aware of their capabilities and the requirements of test items, and the strategic decision-making associated with correct outcomes. RE-based specifications can be used to identify specific cognitive processes embedded in individual item types. Identifying which items individual learners are successful in and which they are not can help to identify reading abilities that learners need to target. Thus, RE can also have a diagnostic function. Teachers can provide cognitive feedback to their students by informing them whether they are using appropriate search strategies, or basing their responses on an appropriate unit of information (word; clause; sentence; multiple sentences; paragraph; text) that is required to answer specific items.

The next section outlines the contribution of the thesis to the methodology used as part of the RE process. This study has demonstrated that evidence-centred design (Mislevy, 2003a, 2003b) offers a clear framework for a research agenda of RE, although the study has also revealed that RE is subject to the limitations of any exploratory framework, and will also be

subject to the same advantages and disadvantages of any methodology adopted for the RE process.

5.6. Methodological innovations and research implications

RE in the present study operates within Mislevy's (2003) conceptual assessment framework (CAF) – assembly, presentation and delivery models. This research agenda has fulfilled the requirements of the definition of critical RE as forwarded by Davidson and Lynch (2002) and Fulcher and Davidson (2007) by providing sufficient details to produce a representative cognitive specification of existing instruments. The use of Toulmin's (2003) interpretive argument to underpin the RE framework led to critical reflection of the methodology of stimulated-recall interviews in an effort to reassess how to link emergent data from test performance to explicit claims about the cognitive processing undertaken by participants in the study. Nonetheless, the RE framework allows for scope to adopt a variety of methodologies that fit within the interpretive argument structure of Toulmin's model of argument. The strength of claims associated with RE will therefore be subject to the limitations of any methodology adopted for the RE.

The research was carefully designed to mitigate the known drawbacks of SRI, such as 'veridicality' (the accuracy of participant claims), 'reactivity' (the influence of the artificial procedure on participant claims), memory decay and completeness. The nature of the video-stimulus and additional participant response stimuli and interview structure allowed participants to formulate and structure their own verbalisations regarding points of significance to them. The video acts as a vivid stimulus of short-term memory, ensuring an objective source of participant actions to increase the likelihood that participants are verbalising their thought processes as they occurred. Video stimulus highlighted another factor associated with SRI which is not discussed in the literature – *relatability* of the stimuli. This refers to the extent to which the stimuli can be easily understood by the participants. Consideration of the extent to which participants can relate the stimuli to their own mental processes during task completion is an important factor in the quality of subsequent verbalisations. Future SRI studies, particularly those which involve eye-tracking as a data collection methodology, which utilise stimuli such as gaze-plots, should consider what

stimuli are best to gather relevant data for their research questions. Gaze plots are fairly opaque stimuli which may produce inadequate verbalisations for specific research purposes. Additionally, SRI should be preferred to questionnaires in future studies of cognitive processing. Adult L2 learners possess well-developed and stable cognitive capabilities, whereas strategic management may change rapidly depending on their familiarity with task types and instructions. The thesis has demonstrated that test takers are clearly aware of their own information processing, even in an artificial and unfamiliar SRI context (Phakiti, 2007).

One possible implication of the methodology is the blurring of the distinction between the automatised or routine metacognitive strategizing related to unconscious actions on the part of the test taker and the overt, metacognitive strategic decisions. Showing participants a video of their performance may cause them to reflect upon their actions and to elaborate what was previously automatized, giving the impression that this was a conscious strategy on the part of the test taker. A further potential drawback is that timing was not an explicit consideration in the current study, meaning that future studies will need to address the outcomes of strictly-timed tests to determine whether the additional time allowed in the present study resulted in more higher-level processes being reported. Finally, it is noteworthy that the methodology required the participants to move through the test in sequence, whereas in a live test, participants may move between items in order to complete those they consider easier. This level of test management was not available to participants, possibly resulting in less authentic procedures being reported. The next section summarises the contribution that the thesis has made towards the reading model of Khalifa and Weir (2009), and therefore the construct of L2 reading more broadly.

5.7. Contribution to the construct of L2 reading

Table 5.1 displays the processing core of Khalifa and Weir's (2009) model of reading as the coding scheme which was applied to the test takers' verbalisations in the SRI. The bottom four layers of Khalifa and Weir's model are presented as 'lower-level processing' and the top five layers as 'higher-level processing'. Two additional coding levels were added to the model. The boundary between lower and higher-level processing in Khalifa and Weir (2009) lies at the sentence boundary. Within sentences, knowledge of grammar, syntax and

vocabulary is sufficient to discern the writer's propositional meaning. Above the sentence level readers are required to link ideas across sentences. Inferential reasoning is cited as a higher-level process in the model, as it implies formulating an idea based on two or more propositions. However, the definition of 'inferencing' offered by Khalifa and Weir includes *cataphoric referencing*. This suggests that the transition between higher and lower-level processing may occur at the *sub-sentence level*, as sentences may contain more than one clause (e.g. relative or subordinate clauses). The addition of two codes at the boundary between lower and higher level processing acknowledges the nuance found at this boundary, by identifying instances of processing which occur at the *clause* level (identifying propositional meaning), the sentence level (in which propositional meaning is gained in a sentence which includes multiple clauses), inferencing at the word level (to account for pronominal referencing occurring within or across sentence boundaries) and inferencing at the clause/sentence level, to account for instances of test takers forming a proposition which is not explicitly stated in the text. Propositional understanding of complex sentences (coded P4s) is arguably a higher level of processing than a basic form of inferential reasoning such as linking pronouns across sentences (coded P5w). Therefore, higher-level processing should be regarded as having commenced at the moment when participants are required to connect ideas *across clauses*. More complex processing can occur within sentences than across them.

The study hypothesised that each of the levels would be able to be applied to participant verbalisations in SRI if participants received sufficient stimuli. This thesis has demonstrated that eight of the nine levels in Khalifa and Weir's (2009) model can be applied to SRI verbalisations, with only the highest level omitted as neither of the tests which were scrutinised test this level of cognitive processing. This study has also expanded upon the definition of inferential reasoning in Khalifa and Weir's model to incorporate both word and sentence-level inferential reasoning, and reconsidered where the boundary lies between higher and lower-level processing in the model.

5.8. Final words and reflection

This thesis is the culmination of a process that began in 2011 and has attempted to bring together some quite disparate elements. 'Reverse engineering' was an idea that was included in some influential works (Davidson and Lynch, 2002; Fulcher and Davidson, 2007) in language testing, yet had received scant attention beyond general discussion. However, the concept of RE was not new in itself, having long associations with systems engineering. Reading this literature, there were striking analogies to language test development. This thesis has brought these two bodies of literature together via evidence-centred design (ECD), a traditionally 'forward engineering' body of literature aimed at developing a test from beginning to completion.

It is clear from the thesis that even within a principled framework of RE, the procedure will be subject to the same strengths and limitations of the chosen methodology. A formal RE framework brings together systems engineering and evidence-centred design within an interpretive framework in order to facilitate the creation of representative specifications which target the cognitive or mental aspects which a language test is designed to reveal. This thesis will form the basis for future papers developing the concept of RE for practitioners in the field. Hopefully, readers are convinced that this is productive and useful concept for the field of language testing.

Appendices

Appendix A

Declaration

Date: / /

I,_____ **[PRINT NAME]**, state that I am over 18 years of age. I voluntarily agree to participate in a research project conducted by Nathaniel Owen, a doctoral candidate studying in the School of Education, University of Leicester.

I understand that this research forms part of doctoral thesis project and is motivated by a desire to understand how test takers interact with the reading components of two tests of English as a foreign language (IELTS and TOEFL). I understand that the specific tasks I will perform require me to complete two tests of English as a foreign language and participate in stimulated-recalls interview. This participation will last approximately 90 minutes. I agree to audio recordings of my verbalisations and video recordings of my test performance. I understand that extracts of my speech will be transcribed for the purposes of analysis.

I understand that no personal data will be retained either by the researcher or the University of Leicester. I acknowledge that the researcher, Nathaniel Owen, has explained the task to me fully; has informed me how I can withdraw from the participation without prejudice or penalty; has offered to answer any questions that I might have concerning the research procedure; has assured me that any information that I give will be used for research purposes only and that this will be treated with the strictest confidence and will remain anonymous and safeguarded, subject to any legal requirements. I understand that some of the transcripts of the test may be used in research documents and may be published in academic journals.

This research does not involve any form of deception, withdrawal of information or misleading information. Without needing to give any reason, at any time during your participation, you are encouraged to feel comfortable to express to the researcher if you wish to withdraw your participation. The researcher will halt the session and no data will be retained. Upon withdrawal any information that you have provided will be securely destroyed and removed from the study. Please be assured that you will not experience any adverse consequences of your withdrawal from the study, as your participation is voluntary. You can withdraw yourself from the study regardless of initial consent during any point of the test.

In discussion of the research, all data emanating from the task will remain anonymous, and any extracts of speech or written material will be sourced anonymously.

I, the participant, understand that by signing to consent to participate in this study does not oblige me to participate. I understand I can withdraw myself from the study at any point.

Thank you for your participation. It is greatly appreciated.

Signature of Participant

.....
Signature of Researcher

.....

References

Abraham, R. and Vann, R. (1996). 'Using Task Products to Assess Second Language Learning Processes'. *Applied Language Learning*, 7 (1-2), 61-89.

Afflerbach, P., Pearson, P. D. and Paris, S. G. (2008). 'Clarifying differences between reading skills and reading strategies'. *The Reading Teacher*, 61 (5), 364–373.

Alderson J. C. (1984). 'Reading in a foreign language: A reading problem or a language problem?' In Alderson J. and Urquhart A. (Eds.), *Reading in a foreign language*, 1–27. Cambridge: Cambridge University Press.

Alderson, J. C. (2000). *Assessing Reading*. Cambridge: Cambridge University Press.

Alderson, J.C. and C. M. Clapham (Eds.) (1992) 'Examining the ELTS Test: An Account of the First Stage of the ELTS Revision Project'. *IELTS Research Report Vol. 2.* Cambridge: The British Council, UCLES and IDP Australian Universities and Colleges.

Alderson, J. C., Clapham, C. and Wall, D. (1995). *Language Test Construction and Evaluation.* Cambridge: Cambridge University Press.

Alderson, J. C., Figueras, N., Kuijper, H., Nold, G., Takala, S. and Tardieu, C. (2004). *The Development of Specifications for Item Development and Classification within the Common European Framework of Reference for Languages: Learning, Teaching, Assessment Reading and Listening.* Final Report of the Dutch CEF Construct Project Available online: http://eprints.lancs.ac.uk/44/1/final_report.pdf (Date accessed 18th November, 2015).

Almond, R.G., Steinberg, L.S., and Mislevy, R.J. (2002). 'Enhancing the design and delivery of assessment systems: A four-process architecture'. *Journal of Technology, Learning, and Assessment*, 1 (5).

Anderson, N.J. (1989). 'Reading Comprehension Tests versus Academic Reading: What are Second Language Readers Doing?' Unpublished doctoral dissertation, University of Texas at Austin.

Anderson, N. J. (1991). Individual differences in strategy use in second language reading and testing. *Modern language Journal*, 75, 460-472.

Anderson, N. J., Bachman, L., Perkins, K., and Cohen, A. (1991). 'An exploratory study into the construct validity of a reading comprehension test: Triangulation of data sources'. *Language Testing, 8*(1), 41-66.

Anderson, R., and W. Nagy. (1991). 'Word meanings', in Barr, R., Kamil, M., Mosenthal, P. and Pearson, P. D. (Eds.), *Handbook of Reading Research*, 2, 690–724. New York: Longman.

Bachman, L. F. (1990) *Fundamental considerations in language testing.* Oxford: Oxford University Press.

Bachman, L. F. (2005). 'Building and supporting a case for test use'. *Language Assessment Quarterly*, 2 (1), 1-34.

Bachman, L. F., Davidson, F., Ryan, K. and Choi, I. C. (1995). *An Investigation into the Comparability of Two Tests of English as a Foreign Language: The Cambridge TOEFL Comparability Study*. Studies in Language testing Volume 1. Cambridge: Cambridge University Press.

Bachman, L. F. and Palmer, A. S. (1996). *Language testing in practice*. Oxford: Oxford University Press.

Bachman, L. F. and Palmer, A. S. (2010). *Language assessment in practice*. Oxford: Oxford University Press.

Bax, S, (2013). 'The cognitive processing of candidates during reading tests: Evidence from eye- tracking' *Language Testing*, 30 (4), 41–465.

Berardi-Coletta, B., Buyer, L. S., Dominowski, R. L., and Rellinger, E. R. (1995). 'Metacognition and problem solving: A process-oriented approach'. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 21(1), 205-223.

Bereiter, C., and Bird, M. (1985). 'Use of thinking aloud in identification and teaching of reading comprehension strategies'. *Cognition and Instruction*, 2, 131–156.

Block, E. (1986). 'The comprehension strategies of second language readers', *TESOL Quarterly*, 20, 463-494.

Bowles, M. A. (2008). 'Task type and reactivity of verbal reports in SLA: A first look at an L2 task other than reading'. Studies in Second Language Acquisition, 30 (4), 359-387.

Bowles, M. (2010). *The think-aloud controversy in second language research*. New York: Routledge. Oxford.

Bowles, M. A. and Leow, R. P. (2005). 'Reactivity and type of verbal report in SLA research methodology'. *Studies in Second Language Acquisition*, 27, 415-440.

Buck, G. (2001). *Assessing Listening*. New York: Cambridge University Press.

Canale, M. and Swain, M. (1980). 'Theoretical bases of communicative approaches to second language teaching and testing', *Applied Linguistics*, 1, 1-47.

Carr, N. (2011). *Designing and Analysing Language Tests: A Hands-on Introduction to Language Testing Theory and Practice*. Oxford: Oxford University Press.

Carrell, P. (1982). 'Cohesion is not Coherence', *TESOL Quarterly*, 16 (4), 479-488.

Chapelle, C. (2012). 'Validity argument for language assessment: The framework is simple…', *Language Testing,* 29: 3, 19-27.

Chapelle, C., Enright, M. K., and Jamieson, J. (2010). 'Does an argument-based approach to validity make a difference?' *Educational Measurement: Issues and Practice*, 29: 1, 3-13.

Charge, N. and Taylor, L. B. (1997) Recent developments in IELTS. *ELT Journal*, 51, 374-380.

Chikofsky, E. J. and Cross, J. H. II. (1990). 'Reverse Engineering and Design Recovery: A Taxonomy', *IEEE Software*, 7(1), 13-17.

Chomsky, Noam. (1965). *Aspects of the Theory of Syntax*. Cambridge, MA: MIT Press.

Clapham, C. (1996). 'The development of IELTS: A study of the effect of background knowledge on reading comprehension'. *Studies in Language Testing Vol. 4*. Cambridge: Cambridge University Press.

Cohen, A. (1994). 'English for academic purposes in Brazil: the use of summary tasks', in C. Hill and K. Parry, (eds.), *From testing to assessment: English as an international language*, 174-204.

Cohen, A. (2006). The coming of age of research on test-taking strategies. *Language Assessment Quarterly*, 3(4), 307-331.

Cohen, A. (2007). The coming of age for research on test-taking strategies, in J. Fox, M. Weshe, D. Bayliss, L. Cheng, C. Turner, and C. Doe (Eds.), *Language testing reconsidered*, 80-111. Ottawa: Ottawa University Press

Cohen, A. and Upton, T. (2006). *Strategies in Responding to the New TOEFL reading tasks*, TOEFL Monograph Series MS-33. Princeton: NJ: Educational Testing Service.

Cohen, A. (2011). *Strategies in learning and using a second language*. Harlow, England: Longman Applied Linguistics/Pearson Education.

Cohen, A. (2012). 'Test-taking Strategies and task design', in Fulcher, G. and Davidson, F. (eds.) *The Routledge Handbook of Language Testing*. New York and London: Routledge, 262-277.

Davidson, F. (2012). 'Releasability of Language Test Specifications', *Japan Language Testing Association (JLTA) Journal* 15: 1-23.

Davidson, F. and Lynch, B. K. (2002) *Testcraft: A Teacher's Guide to Writing and Using Language Test Specifications.* New Haven, CT: Yale University Press.

Davies, A. (2012). 'Kane, Validity and Soundness'. *Language Testing,* 29: 37-42.

Duran, R., Canale, M., Penfield, J., Stansfield, C., and Liskin-Gasparro, J. (1985). *TOEFL from a communicative viewpoint on language proficiency: A working paper* (TOEFL Research Rep. No. 17). Princeton, NJ: ETS.

Educational Testing Service (2009). *The Official Guide to the TOEFL test*. 3rd Edition. Princeton, NJ: ETS.

Educational Testing Service (2015). 'Who accepts test scores?' Available online: https://www.ets.org/toefl/ibt/about/who_accepts_scores (17th December 2015)

El Atia, S. (2004). 'The Baccalauréat exam in France: history, merit, and passing', *Journal of Contemporary French Civilization*, 28 (1), 111-120.

El Atia, S. (2008). 'From Napoleon to Sarkozy: 200 years of the Baccalauréat exam.' *Language Assessment Quarterly,* 5 (2), 142-153.

Ellis, R. (1994). *The Study of Second Language Acquisition*. Oxford, UK: Oxford University Press.

Ellis, R. (2001), Introduction: Investigating Form-Focused Instruction. *Language Learning*, 51: 1–46.

Enright, M. K. and Schedl, M. (2000). *Reading for a reason: Using reader purpose to guide test design*. Unpublished manuscript. ETS.

Enright, M. K., Grabe, W., Koda, K., Mosenthal, P., Mulcahy-Ernt, P. and Schedl, M. (2000). *TOEFL 2000 Reading Framework: A Working Paper*. TOEFL Monograph Series MS-17. Educational Testing Service. Princeton, NJ: ETS.

Enright, M. K., Bridgeman, B., Eignor, D., Kantor, R. N., Mollaun, P., Nissan, S., Powers, D. E., Schedl, M. (2008). 'Prototyping New Assessments', in Chappelle, C., Enright, M. K. and Jamieson, J. M. (eds.) *Building a Validity Argument for the Test of English as a Foreign Language*, New York: Routledge.

Ericsson, K. and Simon, H. (1980). 'Verbal Reports as Data'. *Psychological Review*, 87, 215-251.

Ericsson, K. and Simon, H. (1993). *Protocol Analysis: Verbal Reports as Data* (2nd ed.). Boston: MIT Press.

Eilam, E. (2005). *Reversing: Secrets of Reverse Engineering,* Indianapolis: Wiley Publishing.

Faerch, C. and Kasper, G. (1987). 'From Product to Process – Introspective Methods in Second Language Research', in Faerch, C. and Kasper, G. (Eds.), *Introspection in Second Language Research*, 5-23. Clevedon: Multilingual Matters.

Feldman, U., and Stemmer, B. (1987). 'Thin__aloud a__retrospective da__in Cte__taking: diff__languages__diff__learners__sa__approaches?' In Færch, C. and Kasper,G. (Eds.), Introspection in second language research (5-23). Clevedon: Multilingual Matters.

Field, J. (2004). *Psycholinguistics: the key concepts.* Routledge, London.

Field, J. (2008). *Listening in the language classroom*. Cambridge, UK: Cambridge University Press.

Field, J. (2011). 'Cognitive Validity', in Taylor, L. (Ed.) *Examining Speaking*, Studies in Language Testing, 30, 65-111. Cambridge: Cambridge University Press.

Field, J. (2013). 'Cognitive Validity', in Geranpayeh, A. and Taylor, L. (eds.) (2013). *Examining Listening*, Studies in Language Testing, 35, 77-151. Cambridge: Cambridge University Press.

Freedle, R. (1997). 'The relevance of multiple -choice reading test data in studying expository passage comprehension: The saga of a 15-year effort towards an experimental/correlational merger'. *Discourse Processes*, 23, 399-440.

Freedle, R., and Kostin, I. (1993). 'The prediction of TOEFL reading comprehension item difficulty for expository prose passages for three item types: Main idea, inference, and supporting idea items', *TOEFL Research Report*, 44. Princeton, NJ: Educational Testing Service.

Fulcher, G. (1999). 'Assessment in English for Academic Purposes: Putting Content Validity in its Place', *Applied Linguistics*, 20 (2), 221-236.

Fulcher, G. and Davidson, F. (2007) *Language Testing and Assessment: An Advanced Resource Book* (Routledge Applied Linguistics Series). Oxford: Routledge.

Fulcher G., and Davidson, F. (2009). 'Test architecture, test retrofit'. *Language Testing,* 26 (1), 123–144.

Gass, S. and Mackey, A. (2000). *A. Stimulated Recall Methodology in Second Language research.* Routledge: New York and London.

Glaser, R. (1991). Expertise and assessment, in Wittrock, M. C. and Baker, E. L. (eds.) Testing and Cognition, Englewood Cliffs, Prentice Hall, 17-30.

Gordon, C. (1987). *The effect of testing method on achievement in reading comprehension tests in English as a foreign language*. Unpublished master's thesis, Tel Aviv University, Ramat-Aviv, Israel.

Grabe, W. (2000). 'Reading research and its implications for reading assessment', in A. Kunnan (Ed.), *Fairness and validation in language assessment*, 226-260. Cambridge: Cambridge University Press.

Graesser, A. C., McNamara, D. S., Louwerse, M. M. and Cai, Z. (2004). 'Coh-Metrix: Analysis of text on cohesion and language'. *Behaviour Research Methods, Instruments, and Computers,* 36 (2), 193-202.

Graesser, A. C., Singer, M. and Trabasso, T. (1994). 'Constructing Inferences During Narrative Text Comprehension', *Psychological Review*, 101 (3), 371-395.

Grant, T. (2004), *International Directory of Company Histories*, 62. St. James Press.

Green, A. (2007). *IELTS washback in context: Preparation for academic writing in higher education*. Studies in Language Testing, 25. Cambridge University Press.

Green, A., Khalifa, H. and Weir, C. J. (2013). 'Examining textual features of reading texts - a practical approach'. *Cambridge ESOL Research Notes*, 52, May, 24-39.

Hawkey, R, (2006). *Impact Theory and Practice: Studies of the IELTS test and Progetto Lingue 2000*, Cambridge University Press, Cambridge.

Henning, G., and Cascallar, E. (1992). *A preliminary study of the nature of communicative competence* (TOEFL Research Report 36). Princeton, NJ: ETS.

Hu, M., and Nation, I.S.P. (2000). 'Vocabulary density and reading comprehension'. *Reading in a Foreign Language*, 13 (1), 403–430.

Hymes, D. H. (1972). 'On Communicative Competence', in: J.B. Pride and J. Holmes (Eds.) *Sociolinguistics. Selected Readings*. Harmondsworth: Penguin, 269-293.

IELTS (2007) The IELTS Handbook, Cambridge: Cambridge University Press.

IELTS (2015) Test takers information. Available online: http://www.ielts.org/test_takers_information/question_types/question_types_-_ac_reading.aspx (Date accessed 17th December, 2015).

Jamieson, J. M., Eignor, D., Grabe, W. and Kunnan, A. J. (2008). 'Frameworks for a New TOEFL', in Chappelle, C., Enright, M. K. and Jamieson, J. M. (Eds.) *Building a Validity Argument for the Test of English as a Foreign Language*, New York: Routledge.

Jamieson, J., Jones, S. Kirsch, I., Mosenthal, P. and Taylor, C. (2000). *TOEFL 2000 Framework: A working paper*. TOEFL Monograph Series MS-16. Princeton: Educational Testing Service.

Kane, M. (1992). An argument-based approach to validity. *Psychological Bulletin*, Vol. 112 (3), 527-535.

Kane, M. (2006). Validation, in Brennan, R. (Ed.), *Educational Measurement,* 4th Edition Westport, CT: American Council on Education and Praeger, 17-64.

Khalifa, H., and Weir, C. J. (2009). *Examining Reading: Research and Practice in Assessing Second Language Reading*, Studies in Language Testing volume 29, Cambridge: Cambridge University Press.

Knoblich, G., and Rhenius, D. (1995). Zur Reaktivität Lauten Denkensbeim komplexen Problemlösen[The reactivity of thinking aloud during complex problem-solving]. Zeitschrift für Experimentelle Psy-chologie, 42, 419-454.

Koda, K. (2000). 'Cross-linguistic variations in L2 morphological awareness'. *Applied Psycholinguistics*, 21, 297-320.

Koda, K. (2005). *Insights into second language reading: A cross-linguistic approach.* New York, NY: Cambridge University Press.

Mackey, A. Gass, S. and McDonough, K. (2000). 'How do learners perceive interactional feedback?' *Studies in Second Language Acquisition*, 22, 471-497.

Leow, R. P. and Morgan-Short, K. (2004). 'To think aloud or not to think aloud: The issue of reactivity in SLA research methodology.' *Studies in Second Language Acquisition*, 26, 35-57.

Messick, S. (1989). 'Validity', in Linn, R. L. (ed.) *Educational Measurement*. New York: Macmillan/American Council on Education, 13-103.

Messick, S. (1996). 'Validity of Performance Assessment', in Philips, G. (Ed.) *Technical Issues in Large-Scale Performance Assessment.* Washington, DC: National Center for Educational Statistics.

Mislevy, R. J., Almond, R. G. and Lukas, J. F. (2003). *A Brief Introduction to Evidence-centred Design.* Research Report RR-03-16. Princeton, NJ: Educational Testing Service.

Mislevy, R.J. and Haertel, G.D. (2006). 'Implications of Evidence-centered Design for Educational Testing', *Educational Measurement: Issues and Practice*, 25 (4), 6–20.

Mislevy, R., and Riconscente, M. (2005). 'Evidence-centered assessment design: Layers, structures, and terminology', *PADI Technical Report*, 9. Menlo Park, CA: SRI International.

Mislevy, R., Steinberg, L., and Almond, R. (2000). *Leverage points for improving educational assessment.* A paper prepared for an invitational meeting, entitled The Effectiveness of Educational Technology: Research Design for the Next Decade, Menlo Park, CA, SRI International.

Mislevy, R. and Yin, C. (2012). 'Evidence Centred Design', In Fulcher, G. and Davidson, F. (Eds.). *The Routledge Handbook of Language Testing*. London and New York: Routledge.

Moore, T., Morton, J. and Price, S. (2007). 'Construct validity in the IELTS Academic Reading test: A comparison of reading requirements in IELTS test items and in university study', IELTS Research Reports, Vol. 11 (4).

Morgan, J. L. and Sellner, M. B. (1980). 'Discourse and Linguistic Theory', in Bertram, B. C. and Brewer, W. F. (eds.) *Theoretical Issues in Reading Comprehension.* Hillsdale, NJ: Lawrence Erlbaum Associates.

Muljani, M., Koda, K., and Moates, D. (1998). Development of L2 word recognition: A Connectionist approach. *Applied Psycholinguistics*, 19, 99-114.

Nation, P. and Newton, J. (1997). 'Teaching vocabulary', in Coady, J. and Huckin, T. (Eds.), *Second Language Vocabulary Acquisition. New York*: Cambridge University Press. 238-254.

Nelson, T. O., and Narens, L. (1990). Metamemory: A theoretical framework and new findings, in G. H. Bower (Ed.), The psychology of learning and motivation (1–45). New York: Academic Press.

Nevo, N. (1989). Test-taking strategies on a multiple-choice test of reading comprehension, Language Testing 6, 199-215.

Nyhus, S. E. (1994). *Attitudes of non-native speakers of English toward the use of verbal report to elicit their reading comprehension strategies*. Unpublished M.A. thesis. Minneapolis: University of Minnesota.

Oller, J. W., Jr. (2012). 'Grounding the argument-based framework for validating score uses'. *Language Testing,* 29 (3), 29-36.

O'Sullivan, B. (2011*). Language testing. Theories and practices*. Basingstoke, UK: Palgrave Macmillan.

Oxford, R. L. (ed.) (1996). *Language learning strategies around the world: Cross cultural perspectives*. University of Hawai'i at Manoa: Second language teaching and curriculum centre, Technical Report 13.

Pearlman, M. (2008). 'Finalising the Test Blueprint', in Chapelle, C. A., Enright, M. K. and Jamieson, J. M. (Eds.) *Building a Validity Argument for the Test of English as a Foreign Language.* New York: Routledge, 227-258.

Perea L. (2011). 'Benefits of teachers' feedback to reverse-engineering item language test specifications from an existing item bank', *Texas Papers in Foreign Language Education*. 15 (1), 30-54.

Phakiti A (2003). 'A closer look at the relationship of cognitive and metacognitive strategy use to EFL reading achievement test performance'. *Language Testing* 20, 26-56.

Phakiti A (2008). 'Construct validation of Bachman and Palmer's (1996) strategic competence model over time in EFL reading tests'. *Language Testing* 25, 237-72.

Popham, W. J. (1978) *Criterion-referenced measurement.* Englewood Cliffs, NJ: Prentice Hall.

Pressley, M., and Afflerbach, P. (1995). *Verbal protocols of reading: The nature of constructively responsive reading.* Hillsdale NJ: Erlbaum.

Purpura, J. E. (1999). *Learner strategy use and performance on language tests: A structural equation modeling approach*. Cambridge: Cambridge University Press.

Rasch, G (1960). *Probabilistic Models for Some Intelligence and Attainment Tests*. Copenhagen: Danish Institute for Educational Research. (Expanded edition, 1980). Chicago: University of Chicago Press.

Rekoff Jr., M. G. (1985). 'On Reverse Engineering', *IEEE Trans. Systems, Man and Cybernetics,* 244 – 252.

Selinker, L. (1972), 'Interlanguage', *International Review of Applied Linguistics*, 10, 209–241.

Singer, M. (2007). 'Inference processing in discourse comprehension', *The Oxford Handbook of Psycholinguistics*, Gaskell, G. (Ed.) Oxford: Oxford University Press, 343 – 360.

Stemmer. B. (1991). *What's on a C-test taker's mind? Mental processes in C-test taking*. Bochum: Universitätsverlag Brockmeyer.

Swales, J.M. (1990) *Genre Analysis: English in Academic and Research Settings.* Cambridge: Cambridge University Press.

Shaw, S.D. and Weir, C. J. (2007). *Examining Writing: Research and practice in Assessing Second Language Writing*. Studies in Language Testing 26, Cambridge: Cambridge University Press.

Toulmin, S. E. (2003). *The Uses of Argument.* (Updated edition). Cambridge: Cambridge University Press.

Urquhart, S. and C. Weir (1998). *Reading in a Second Language: Process, Product and Practice*. London: Addison Wesley Longman Ltd.

Vigliocco, G. and Vinson, D. P. (2007). 'Semantic Representation', In Gaskell, G. (Ed.) *Handbook of Psycholinguistics*. Oxford: Oxford University Press.

Warren, J. (1996) 'How students pick the right answer: A 'think aloud' study of the French CAT', in J. Burston, M. Monville-Burston, and J. Warren (Eds.), *Issues and innovations in the teaching of French*, Occasional paper No. 15, 79-94. Australian Review of Applied Linguistics.

Watson, J. (1913). 'Psychology as the behaviourist views it'. *Psychological Review*, 20, 158-177.

Watson, J. (1920). 'Is thinking merely the action of language mechanisms?' *British Journal of Psychology*, 11, 87-104.

Weir, C. J. (2005). *Language testing and validation: An evidence-based approach,* Palgrave MacMillan, Basingstoke.

Weir, C. J, Hawkey, R., Green, A., Devi, S. and Unaldi, A., (2009a). 'The relationship between the academic reading construct as measured by IELTS and the reading experiences of students in their first year of study at a British university', in *Research Reports*, Volume 9, Thompson, P. (Ed.) British Council/IDP Australia, London.

Weir, C.J., Hawkey, R., Green, A. and Devi, S. (2009b), 'The cognitive processes underlying the academic reading construct as measured by IELTS', in *Research Reports*, Volume 9, Thompson, P. (Ed.) British Council/IDP Australia, London.

Yin, R. (2010). 'Analytic Generalization', In Albert J. Mills, G. Durepos, and E. Wiebe (Eds.), *Encyclopaedia of Case Study Research*. (21-23). Thousand Oaks, CA: SAGE Publications, Inc.