Toward a Cognitive Processing Model of MELAB Reading Test Item Performance

Lingyun Gao
University of Alberta

This study develops and tests a model of cognitive processes hypothesized to underlie MELAB reading item performance. The analyses were performed using the reading section on two MELAB forms using a three-pronged procedure: (1) reviewing theoretical models of L2 reading processes and constructs of L2 reading ability, (2) analyzing cognitive demands of the MELAB reading items and collecting verbal reports of the cognitive processes actually used by examinees when correctly answering the items, and (3) examining the relationship between the proposed cognitive processes and empirical indicators of item difficulty using a cognitively based measurement model, the tree-based regression (TBR). While the results were inconsistent across forms, the processes of drawing inferences and evaluating alternative options accounted for a significant amount of the variance in MELAB reading item difficulty on the two forms. Results of this study inform the construct validation of the MELAB reading and item construction, and lay a foundation for the MELAB reading as a diagnostic measure.

Large-scale assessments are widely used for a variety of purposes such as admissions, matching students to appropriate instructional programs, and enhancing learning (National Research Council [NRC], 2001). Assessment results typically provide a percentile rank to reveal where an examinee stands relative to others, or a numeric score to indicate how the examinee has performed. The one challenge with most large-scale assessments, however, is the lack of capacity to interpret more complex forms of evidence derived from examinees' performance (Embretson & Gorin, 2001; NRC, 2001). Consequently, assessments provide very limited information to test developers and users, the validity of the inferences drawn from the assessment results is frequently questioned, and the usefulness of the assessments as a learning tool is compromised (Alderson, 2005a; Gorin, 2002; Strong-Krause, 2001).

In the last decade, within the language testing and measurement communities, there has been a growing emphasis on the union of cognitive psychology and assessments to yield meaningful information regarding examinees' knowledge structure, skills, and strategies used during task solving (Cohen & Upton, 2005; Douglas & Hegelheimer, 2005; Embretson, 1999; Leighton, 2004; Mislevy, 1996; Mislevy, Steinberg, & Almond, 2002; Nichols, 1994; NRC, 2001). One approach to achieving this goal has been to model item statistical properties, in particular item difficulty, in terms of the cognitive processes involved in item solving (Embretson, 1998; Huff, 2003). To date, a number of models have been developed linking item statistics to item features for a variety of foreign/second language assessments (e.g., Carr, 2003; Kostin, 2004). However, only a few models have linked item statistics to the cognitive structure of test items (e.g., Sheehan & Ginther, 2001), and many of these models are limited by the concepts and methods employed. Conceptually, due to the gaps among cognitive psychology, measurement, and subject areas, many of the existing models fail to incorporate the most current cognitive theories in a particular domain, which are critical for defining item features and interpreting the models. Methodologically, some of the most useful methods that cognitive psychologies use to understand human thought, such as task analysis, protocol analysis, and the study of reaction times, have not been widely used to explain test item performance. In addition, due to technical complexity, advanced measurement models incorporating cognitive elements, such as the rule-space model (Tatsuoka, 1995), tree-based regression (Sheehan, 1997), and Bayes inference networks (Mislevy, Almond, Yan, & Steinberg, 1999) have not been widely applied to assessment practice. Much work is required to link critical features of cognitive models specific to a substantive testing context to new measurement models and to observations that reveal meaningful cognitive processes in a particular domain (NRC, 2001).

The Michigan English Language Assessment Battery (MELAB) is developed by the English Language Institute at the University of Michigan (ELI-UM) to assess the advanced-level English language competence of adult nonnative speakers of English who will use English for academic study in an American university setting. The MELAB is used primarily for higher education admission, and the assessment results are widely accepted as evidence of English competence by educational institutions in the countries where English is the language of instruction. The MELAB consists of Part 1, composition, Part 2, a listening test containing 50 multiple-choice items, Part 3, a grammar/cloze/vocabulary/reading test containing a total of 100 multiple-choice items, and an optional speaking test. Compositions and speaking tests are scored by trained raters using rating scales. Answer sheets for Part 2 and Part 3 are computer scanned and raw scores are converted to scale scores. The MELAB reports a score for each part and the final score, which is the average of the scores on the three parts.

Current MELAB score reporting provides some information on examinees' English competence and describes examinees' competence in writing and speaking to some extent. However, a numeric score for Part 2 and Part 3 provides limited information to examinees, admissions officers, and other stakeholders regarding examinees' strengths and weaknesses in listening and especially in reading, where a subscore is lacking. Reading is a major part of language acquisition and language use activity in everyday life (Grabe & Stoller, 2002). In the context of using English as a second or foreign language for academic purposes, reading tends to be the single most important language skill and language use activity that nonnative English speakers need for academic activities (Carr, 2003; Cheng, 2003). Hence, the nature of reading in a second or foreign language and how to assess it on large-scale high-stakes tests have been a primary concern for language researchers and testers (Alderson, 2000; 2005a; 2005b; Bernhardt, 2003; Cohen & Upton, 2005).

The purpose of this study is to model the cognitive processes underlying performance on the reading items included in the MELAB using a cognitive-psychometric approach. The specific research questions (RQ) are:

1. What cognitive processes are required to correctly answer the MELAB reading items?
2. What cognitive processes are actually used by examinees when they correctly answer the MELAB reading items? How are they related to the findings in response to RQ 1?
3. To what extent do the cognitive processes used to correctly answer the MELAB reading items explain the empirical indicators of item difficulty?

Conceptualizing the Cognitive Processes Involved in L2 Academic Reading

Information-Processing Perspectives on Reading

Over the last couple of decades, the shift in psychology from a behavioral to a cognitive orientation has impacted enormously the understanding of reading. Bottom-up processing is an immediate left-to-right processing of the input data through a series of discrete stages (Ruddell, Ruddell, & Singer, 1994). Early theories viewed reading as bottom-up processing in which a reader passively and sequentially decoded meanings from letters, words, and sentences (e.g., Anderson, 1972; LaBerge & Samuels, 1974). Reading processes were considered to be completely under the control of the text and had little to do with the information possessed by a reader or the context of discourse (Perfetti, 1995).

Top-down processing is information processing in which readers approach the text with existing knowledge, and work down to the text (Hudson, 1998). The top-down view of reading emphasizes readers' contributions over the incoming textual information. Two representative examples of top-down processing are psycholinguistic models (e.g., Goodman, 1967, Smith, 1971) and schema-theoretic models (e.g., Carrell, 1983a; 1983b). Psycholinguistic models stress the interaction between language and thought, especially readers' inferential abilities, and describe reading processes as active, purposeful, and selective (Smith, 2004). Schema-theoretic models describe the reading process through the activation of schemata (i.e., networks of information organized in memory) and stress the centrality of readers' language and content knowledge. While reading, readers apply their schemata to the text, confirm and disconfirm, and map the incoming information from the printed text onto their previously formed knowledge structures to create meaning (Hudson, 1998). Schema theory is valued at attempting to explain the integration of the new information with old, but fails to explain how completely new information is processed (Alderson, 2000). Critics of schema theory point out that it does not lead to an explicit account of reading processes due to a vague definition of schema, elision of readers' intentionality, and oversimplification of the memory retrieval and storage processes (Phillips & Norris, 2002). Carver (1992) argues that schema theory applies only when reading texts are relatively hard, such as the situation where college-level students study academic texts.

More recent theories of reading stress the simultaneous interaction between bottom-up and top-down processing (e.g., Johnston, 1984; Rumelhart, 1977; 1980; Stanovich, 1980; 2000). According to the interactive theories, readers' multiple sources of knowledge (e.g., linguistic knowledge and world knowledge) interact continuously and simultaneously with text. Current reading theories acknowledge the interactive nature of processing, and emphasize the importance of purpose and context to fluent reading (e.g., Alderson, 2000; Enright, Grabe, Koda, Mosenthal, Mulcahy-Ernt, & Schedl, 2000; Hudson, 1998). As Butcher and Kintsch (2003) note, reading is the interaction among a variety of top-down and bottom-up processes, where readers' knowledge, cognitive skills, strategy use, and purpose of reading are crucial during the process of reading and must be taken into account when modeling text processing.

Models of the L2 Reading Processes

The conceptualization of reading has been evolving over the years, so have the models of the L2 reading processes. Current models of the L2 reading processes generally include language knowledge, background/topical knowledge, cognitive skills, and cognitive and metacognitive strategies. Language knowledge consists of a number of relatively independent components, such as the knowledge of phonology, vocabulary, syntax, and text structure. Major components in current models of the L2 reading processes are discussed below.

Word Recognition

Word recognition has been considered central to fluent reading in current models of reading processes of skilled adult L2 readers (e.g., Alderson, 2000; Grabe, 2002; Hudson, 1996; Koda, 2005; Urquhart & Weir, 1998). It is the process of recognizing strings of letters in print and being able to rapidly identify meanings from visual input (Rayner & Pollatsek, 1989). Unlike skilled adult L1 readers who are generally assumed to have phonological access to the lexicon and are familiar with the script, L2 readers encounter words that they have not heard pronounced and scripts that they are not familiar with in many cases (Urquhart & Weir, 1998). Hence, L2 readers are expected to experience greater difficulty in processing letters in a word and identifying word meanings (Alderson, 2000). In addition, unlike skilled adult L1 readers for whom the words encountered are normally in their lexicon, L2 readers have to handle unfamiliar vocabulary (Urquhart & Weir, 1998). In the context of academic reading, where large amounts of academic texts need to be processed, recognizing words and word meaning is extremely important. Inefficient word recognition and insufficient vocabulary would likely result in inefficient academic reading (Hudson, 1996).

Knowledge of Syntax

Readers must process syntax to impose meaning on the recognized words (Urquhart & Weir, 1998). Syntax is the component of a grammar that determines the way in which words are combined to form phrases and sentences (Radford, 2004). In L2 reading, syntactic knowledge is crucial for successful text processing and has been included in many models of the L2 reading processes (e.g., Alderson, 2000; Grabe, 1991; Hudson, 1996; Koda, 2005).

Knowledge of Textual Features

Readers' knowledge of textual features, such as cohesion and text structure, has long been considered important in text processing (Alderson, 2000; Koda, 2005) and critical to successful L2 academic reading (Hudson, 1996). Cohesion refers to "the connections between sentences," which are furnished by pronouns that have antecedents in previous sentences, adverbial connections, known information, and knowledge shared by the reader (Kolln, 1999, p. 271). Frequently used cohesive devices include reference, substitution, ellipsis, and conjunction. According to Thompson (2004), reference is the set of grammatical resources used to repeat something mentioned in the previous text (e.g., the pronoun "it") or signal something not yet mentioned in the text (e.g., the nondefinite article "A" in the sentences "They came again into their bedroom. *A* large bed had been left in it"). Substitution refers to the use of a linguistic token to replace the repetitive wording (e.g., "I think *so*"). Ellipsis is the set of grammatical resources used to avoid the repetition of a previous clause (e.g., "How old is he? Two years old"). Conjunction refers to the combination of any two textual elements into a coherent unit signaled by conjunctives (e.g., however, by the way, thus). Research has shown that coherent texts contribute to understanding, while ambiguous references, indistinct relationships between elements of the text, and the inclusion of irrelevant ideas or events hinder comprehension (Hudson, 1996; McKeown, Beck, Sinatra, & Losterman, 1992).

Coherence of a text depends on not only cohesive devices but also text structure and organization patterns; that is, how the sentences and paragraphs relate to each other and "how the relationships between ideas are signaled or not signaled" (Alderson, 2000, p. 67). Example text structures include cause/effect, general/specific, problem/solution, comparison/contrast, and the use of definitions, illustrations, classifications, and topic sentences. Research has shown that the internal logic of text structures (strong or weak), organized patterns (tight or loose), and location of information within the text structures (earlier or later) affect processing and understanding (e.g., Carrell, 1984, 1985; Roller, 1990; Hudson, 1996).

Background Knowledge and Subject Matter/ Topic Knowledge

In addition to knowledge of language, readers' background knowledge (i.e., knowledge that may or may not be relevant to the text content) and subject matter/topic knowledge (i.e., knowledge directly relevant to the text content) play a crucial role in L2 reading, especially in L2 academic reading where the reading materials are relatively difficult and the primary concern is to predict examinees' performance on reading tasks involved in academic study (Alderson, 2000; Urquhart & Weir, 1998). According to schema theory and the interactive notion of reading, readers apply their preexisting knowledge when processing texts, which influences the process in which new information is recognized and stored and affects text understanding. Background and topical knowledge have been included in many models of the L2 reading processes (e.g., Grabe, 2002; Hudson, 1996).

Cognitive Skills

In addition to knowledge, readers have skills to learn and process new information in the text. Cognitive skills have long been held as important components of the reading process. For example, Thorndike (1917) stated that reading was reasoning. He explained that readers' skills to construct meaning approximated logical inference and deduction, and that good readers thought clearly. Cognitive skills enable L2 readers to use information in their mind and cues from the text to fill the gap of understanding and monitor their reading processes (Alderson, 2000). Over the last several decades, skills have been a major area in reading research and various taxonomies of L2 reading skills have been developed (e.g., Carver, 1992; Farhady & Hessamy, 2005; Grabe, 1991; Koda, 1996; Munby, 1978). These skill taxonomies provide a framework for reading test construction. However, critics argue that the skills in many of these taxonomies are ill defined, have enormous overlap with one another, and lack empirical support (Alderson, 2000). Despite the criticisms, skills such as inference, synthesis, and evaluation are frequently included in current models of L2 reading processes (e.g., Enright et al., 2000; Hudson, 1996; Jamieson, Jones, Kirsch, Mosenthal, & Taylor, 2000).

Problem-Solving Strategies

In recent L2 reading literature, the strategies used by readers when processing text have received considerable attention (e.g., Abbott, 2005; Cohen & Upton, 2005; Lumley & Brown, 2004; Phakiti, 2003; Yang, 2000). A list of cognitive and metacognitive strategies that L2 readers use during reading include skimming and scanning the text to locate discrete pieces of information, monitoring progress of understanding, planning ahead how to read, selectively attending to text, and, in testing situations, testwiseness strategies (e.g., guessing and attending to the length of options). Cognitive and metacognitive strategies have been important components in many models of the L2 reading processes (Alderson, 2000; Hudson, 1996; Koda, 2005).

Purpose and Context

In addition to knowledge, skills, and strategies, reader purpose and the context in which L2 readers engage in reading is increasingly being emphasized (e.g., Alderson, 2000; Cohen & Upton, 2005; Enright et al., 2000; Hudson, 1996). These researchers stress that reading is usually undertaken for some purpose and in a specific context, which affects the knowledge and skills required, strategies used, and the understanding and recall of the text.

Conceptualizing L2 Academic Reading Ability

The Constructs of L2 Academic Reading Ability

Models of the L2 reading processes suggest a range of constructs of L2 reading ability, which has been operationalized differently in tests of L2 academic reading (e.g., Cohen & Upton, 2005; Douglas, 2000; Enright et al., 2000; Hudson, 1996; Jamieson et al., 2000). It is currently well accepted that word recognition skills, which are critical to fluent reading, need to be tested. Language knowledge is essential for L2 readers' understanding of academic texts, and hence should be measured. Knowledge of formal discourse structure should be taken into account in testing L2 academic reading. Cognitive skills and cognitive and metacognitive strategies are important for L2 readers to overcome the language difficulties, especially when reading difficult academic texts. Hence, L2 academic reading tests should allow examinees to apply their cognitive skills and strategies. Alderson (2000) stresses that in the context of L2 reading, sufficient knowledge of the second or foreign language, cognitive skills, and problem-solving strategies are especially important. Nevertheless, Alderson reminds us that readers' background knowledge is normally not included in the constructs to be assessed, though its influence on the L2 reading process and product is recognized.

A Theoretical Framework of Communicative Language Competence

According to the *MELAB Technical Manual* (English Language Institute, 2003), the framework for developing the MELAB is closely related to the framework of communicative language ability (CLA) proposed by Bachman (1990) and later revised by Bachman and Palmer (1996). Bachman (1990) proposed the framework of CLA, which acknowledges the competence in the language and the capacity for using this competence in contextualized language use (see Figures 4.1 and 4.2, pp. 85–87). Specifically, Bachman's framework of CLA includes language competence, strategic competence, and psychophysiological mechanisms, and describes the interactions of these components with the language user's knowledge structures and language use context. Language competence includes organizational competence, which consists of grammatical and textual competence, and pragmatic competence, which consists of illocutionary and sociolinguistic competence. Strategic competence performs "assessment, planning, and execution functions in determining the most effective means of achieving a communicative goal" (p. 107). Psychophysiological mechanisms are "the channel (auditory, visual) and mode (receptive, productive) in which competence is implemented" (p. 108).

Bachman and Palmer (1996) extended Bachman's (1990) framework and clearly defined language use as the dynamic creation of intended meanings in discourse by an individual in a particular situation. According to Bachman and Palmer, purpose and context of language use are crucial in defining language ability. They stress that to make inferences about language ability based on language test performance, language ability should be defined in a way that is appropriate for a particular testing situation, a particular group of examinees, and a specific context in which examinees will be using the language outside of the test itself. For this study, language use occurs in the context where English competence of adult nonnative English speakers is assessed for academic purposes. Correspondingly, the reading ability includes the language knowledge and strategic competence to solve the test tasks, and the competence to apply the knowledge/competence to academic reading in the real world.

In addition to the emphasis of purpose and context, Bachman and Palmer point out that language use involves complex interactions among individual characteristics of language users and the interactions among these characteristics and the characteristics of language use. Hence, "language ability must be considered within an interactional framework of language use" (pp. 61–62). They presented their framework as a theory of factors affecting performance on language tests and proposed that performance on language tests was affected by (1) the interactions among examinees' language knowledge, topical knowledge, affective schemata, strategic competence or metacognitive strategies, and personal characteristics such as age and native language, and (2) interactions between examinee characteristics and characteristics of the language use, namely, test task. Subsequently, Bachman (2002) clearly distinguished three sets of factors that affected test performance: examinee attributes, task characteristics, and the interactions between examinee and task characteristics. The current theoretical framework of CLA (Bachman, 1990; Bachman & Palmer, 1996) is consistent with current understanding of L2 reading ability and its assessment, which acknowledges the interactive nature of reading and the effect of text and item characteristics, reader knowledge, cognitive and metacognitive strategies, and purpose and context of reading on reading test performance (Alderson, 2000; Enright et al., 2000; Hudson, 1996; Jamieson et al., 2000; Koda, 2005).

Research into L2 Reading Test Item Performance

Methods and Issues Concerning This Research

Over the last decade, language testers have been researching item performance in reading tests. This research has yielded a number of factors that appear to affect item performance across a variety of reading tests. However, limited by the concepts and methods employed, little progress has been made on our understanding of L2 reading test item performance (Bachman, 2002). Conceptually, current theories of reading recognize the interactions between reader and text and emphasize purpose and context of reading (Alderson, 2000). Moreover, current theories of language testing consider task performance as a function of interactions between examinee attributes and test task characteristics (Bachman, 2002). However, many studies of L2 reading test item performance either focus on the characteristics inherent in the text and/or item itself without taking examinees into account, or vice versa. In addition, the varying purposes and contexts of reading tasks were not given proper attention. Methodologically, many of the studies are limited in the analyses employed. Representative studies of L2 reading test item performance are critically reviewed below.

Studies of Surface Task Feature and Item Performance

Studies of surface task features and item performance typically identify a number of text and/or item features and then investigate the effect of these features on item statistics using quantitative methods, such as the commonly used multiple linear regression (e.g., Freedle & Kostin, 1993; 1999). The findings of these studies have clear implications for the design of L2 reading tests. However, due to overreliance on surface features of texts and items without taking readers into account, the analyses fail to reveal the processes of item solving. In addition, multiple linear regression analysis has its limitations, such as oversensitivity to the presence or absence of an item feature variable (Kasai, 1997) and strict requirements for linearity and the number of items (Keppel & Zedeck, 2001).

Freedle and Kostin (1993) examined the effect of task features on the difficulty of TOEFL reading items, as measured by equated delta (n_{items} = 213; $n_{examinees}$ = 2000). Based on a review of previous studies predicting the difficulty of multiple-choice reading test items, they hypothesized that 12 categories of 65 text, item, and text-by-item interaction variables might influence the difficulty of TOEFL reading items. After a multiple-regression analysis, they found that 58% of the variance in item difficulty was explained by eight categories of text and text-by-item variables: negations, referentials, rhetorical organizers, sentence length, passage length, paragraph length, lexical overlap between text and options, and location of relevant text information. Their investigation of reading item difficulty as a function of text, item, and text-item interaction impacted later research and their findings have direct implication for text writing and item design. However, the variables used, which were mainly word counts (e.g., the number of words in the correct option), fail to reveal the complex processes of item solving and lack interpretive and diagnostic value (Kasai, 1997).

Carr (2003) examined task features in explaining the difficulty of 146 reading items included in three TOEFL test forms. Based on a review of previous research, he developed a rating instrument consisting of three sets of 311 passage, key sentence, and item variables, most of which were word and sentence counts. He asked five graduate students in applied linguistics to rate the task features using the rating instrument. However, only passage and key sentence variables were included in his analysis, as text features were considered most relevant to fluent reading and most reflective of the target language use domain. Through exploratory and confirmatory factor analyses, he constructed and tested a factor model of the text features and concluded that passage content, syntactic features of key sentences, and vocabulary factors contributed to the difficulty of the TOEFL reading items. Carr provides a thorough list of text variables that may affect the difficulty of L2 reading test items and an alternative method for investigating the effect of task features on reading item difficulty. However, excluding item variables from the analysis does not seem to be warranted, since the complete task of multiple-choice reading tests involves text, question stem, and options, and examinees' mental processes used to answer multiple-choice items may differ from those used to answer constructed response or essay questions (Kasai, 1997). In addition, like Freedle and Kostin's (1993) study, a focus on the surface features of text provided limited information regarding examinees' cognitive processes during item solving.

Studies of Cognitive Demands of Test Items and Item Performance

Studies of cognitive demands of test items and item performance typically identify item features that are essentially cognitive demands hypothesized to affect the performance of a given item (e.g., Alderson, 1990; Alderson & Lukmani, 1989; Bachman, Davidson, & Milanovic, 1996; Bachman, Davidson, Ryan, & Choi, 1995). These studies used "expert" ratings of the test items that included different combinations of the cognitive demands, and then related the ratings to item performance using cross-table or multiple linear regression analysis. "Experts" in these studies have included various individuals such as EFL teachers or administrators and graduate students in applied linguistics or educational psychology. Results of these studies consistently indicate no systematic relationship between “expert” ratings and item statistics. The equivocal results are likely caused by methodological limitations. For example, no measurement models that incorporate the cognitive elements of items were used in relating the ratings and item statistics. In addition, item statistics calculated using the classical test theory model have little connection with the cognitive structure of an item (Embretson, 1999). Finally, experts may process the test tasks differently from the target examinees (Alderson, 2000; 2005a; Leighton & Gierl, 2005). Nevertheless, these studies begin to pay attention to the effect of cognitive elements of test items on item performance, which anticipates the cognitive processes used by examinees when they answer test items and precludes the study of item performance in light of examinees’ cognitive processes. Expert analysis may reveal both automatic and controlled processes evoked by test items (Leighton, 2004). As automatic processes are inaccessible for description through conscious verbal reports (Cheng, 2003), analysis of the cognitive demands of an item provides valuable sources of data to supplement verbal reports.

Alderson and Lukmani (1989) investigated the cognitive skills required for correctly answering the reading items included in a L2 communication skills test taken by 100 students at Bombay University (India), and related the skill requirements of individual items to item difficulties, as measured by percentage of correct responses. Nine teachers at Lancaster University (Great Britain) were asked to describe what skills were being tested by each of the 41 test items. Results showed little agreement among the judges and little relationship between item difficulty and skill requirements of each item. The lack of a prestructured rating guide and pretraining of the judges is a likely reason for the results. In addition, the judges may not have been familiar with how students processed the test task.

Using a rating instrument containing 14 reading skills, Alderson (1990) conducted a similar study, in which 18 teachers of ESL were asked to decide a single skill being tested by each of 15 short answer questions on two British language proficiency tests. Again, little agreement was reached among the judges and little relationship was found between item difficulty and skill requirements of the items. Two likely reasons for the results might be: (1) correctly answering an item may require multiple skills, while the judges were allowed to specify one skill for each item, and (2) there was enormous overlap among the skills provided on the rating instrument, which affected the accuracy of expert rating.

The studies reviewed above question the ability of experts to determine the skills being tested by an item. Other studies have reported high levels of agreement among expert judges by using well-designed and clearly defined rating instruments, extensive discussion, and exemplification (e.g., Bachman et al., 1995; 1996; Carr, 2003). In Bachman et al.’s (1996) study, five trained applied linguists with experience as EFL teachers were asked to analyze the characteristics of 25 vocabulary and 15 reading items and passages on each of the six parallel forms of an EFL test. The number of examinees for each form ranged from 431 to 1099. A refined rating instrument was presented to the raters, which contained 23 test task features (TTF) and 13 communicative language abilities (CLA) defined using Bachman’s (1990) framework. Rater agreement was checked using generalizability analysis and rater agreement proportion. Results showed that the overall rater agreement was very high and that the TTF ratings were more consistent than the CLA ratings. They related the TTF and CLA ratings to the IRT item parameter estimates calibrated using the 2PL model. Stepwise regressions were performed for all items and for vocabulary and reading items separately, by individual form, and for all forms combined. Results showed that neither TTC nor CLA ratings consistently predicted item difficulty or discrimination across the six forms, though a combination of the TTF and CLA ratings consistently yielded high predictions. Their study demonstrates the possibility of achieving a high level of agreement among expert judges. The use of a rating instrument and rater training appear to play an important part in rater agreement. Their study has several implications. First, more refined definitions of the abilities may increase the consistency of ability ratings. Second, the inconsistent prediction of item parameter estimates across the forms indicates that the item features identified are likely affected by differences among passages and items on different tests. A large number of tests may be examined to provide reliable item features that affect item performance. Finally, as experts may process test tasks differently from the target examinees, it is imperative to examine examinees' actual processes underlying the correct responses (Alderson, 2000; 2005a; Leighton & Gierl, 2005).

Processes in Task Performance Inferred from Verbal Reports

Concurrent verbal reports (i.e., an individual's description of the processes he/she is using during task solving) and retrospective verbal reports (i.e., the recollection of how the task was solved) have been established as valid means to obtain valuable sources of data on cognitive processing during task performance (Ericsson & Simon, 1993). Leighton (2004) recommends collecting both forms of verbal reports to triangulate the processes used to solve the tasks, using tasks of moderate difficulty to maximize the verbalization elicited, and analyzing a task's cognitive demands prior to eliciting verbal reports to anticipate the cognitive processes a respondent will use when solving the task. The last decade has seen an increased use of verbal reports to inspect the processes of L2 readers during test taking (e.g., Abbott, 2005; Allan, 1992; Anderson, Bachman, Perkins, & Cohen, 1991; Block, 1992; Cohen & Upton, 2005; Lumley & Brown, 2004; Phakiti, 2003; Yang, 2000). These studies shed some light on the cognitive processes underlying reading test item performance and suggest a number of processes that appear to predict item statistics. However, as the test tasks differ across the studies, the research results as a whole have been inconsistent.

Anderson et al. (1991) investigated the strategies used by adult EFL learners to complete a standardized reading test and then examined the relationships among strategies, item type, and item performance using a triangulation of three sources of data: retrospective verbal reports, item type, and item difficulty p and discrimination r_{pbi} through chi-square analyses. Their results revealed a significant relationship between (1) frequencies of the reported strategies and item type, and (2) item difficulty and the strategies of skimming, paraphrasing, guessing, responding affectively to text, selecting answer through elimination, matching stem with text, selecting answer because stated in text, selecting answer based on understanding text, and making reference to time. In addition, their results showed that more strategies were reported for the items of average difficulty ($0.33 \leq p \leq 0.67$) than for the difficult items ($p < 0.33$) and easy items ($p > 0.67$). This finding appears to support the use of moderately difficult items to maximize verbal report data (Leighton, 2004). However, no significant relationship was discerned between item type and item difficulty. Their study demonstrates a triangulation approach to the construct validation of a standardized reading test. The authors recommend the use of multiple data sources and stress supplementing the traditional psychometric approach with qualitative analysis of item content and verbal reports, which has significant implications for research on standardized reading tests.

Item Modeling with New Concepts and Methods

In response to the call for the union of cognitive psychology and assessment, there is a growing interest in modeling reading test item performance in light of the cognitive elements of an item in recent psychometric literature (e.g., Gorin, 2002; Huff, 2003; Rupp, Garcia, & Jamieson, 2001; Sheehan 1997; Sheehan & Ginther, 2001). These studies typically rely on expert identification of cognitively based item features, and then relate these features to item statistics using new measurement models that can incorporate these features. These studies have demonstrated that linking "indicator variables that distinguish the cognitive processes assumed to be involved in item solving" and "observable item performance indices, in particular, item difficulties" can provide invaluable validity information and rich sources of data for understanding the cognitive processing during task performance (Wainer, Sheehan, & Wang, 2000, p.114). However, there are several limitations with some of these studies. First, item features are simply judged by experts without being validated by examinees' actual processes while answering items. As experts may process a task differently from the target examinees, expert judgment may not represent examinees' actual processes underlying item performance. Second, item parameter estimates calibrated using the 2-PL or 3-PL IRT models are problematic in the case of passage-based testlets. This is because the interrelatedness among the set of items based on a common passage violates the local item independence assumption of IRT, which can cause inaccurate estimation of examinee abilities and item statistics (Lee, 2004; Kolen & Brennan, 2004; Wainer & Lukhele, 1997). Third, due to the gap between cognitive psychology, measurement, and reading, many of these studies fail to incorporate the most current cognitive theories in reading and fail to justify the item features within a framework of ability constructs. Despite the limitations, psychometric studies on modeling reading item performance with cognitively based item features and tree-based regression (TBR) measurement models offer considerable promise for research into the L2 reading test item performance.

Sheehan (1997) conducted one of the first studies modeling item difficulty based on item processing features in order to develop student- and group-level diagnostic feedback. He analyzed examinee responses to 78 verbal items (40 passage-based reading, 19 analogy, and 19 sentence-completion items) on an operational form of the SAT I Verbal Reasoning Test. In his TBR analysis, the criterion was the 3-PL IRT item difficulty estimates, and the predictors were hypothesized skills required for item solution. Using a user-specified split, the items were first classified according to four processing strategies: Vocabulary, Main Idea and Explicit Statement, Inference, and Application or Extrapolation. The first split explained 20% of the observed variance in item difficulty. To explain more variance, each strategy node was split into two child nodes based on different skills within each strategy. For example, the Vocabulary strategy was further divided into Standard Word Usage and Poetic/Unusual Word Usage. This split explained about 50% more of the observed variance in item difficulty.

In a subsequent study, Sheehan and Ginther (2001) successfully applied TBR to develop an item difficulty model for the Main Idea type reading items on the TOEFL 2000, based on cognitive processing features of the items. They coded the Main Idea items with three variables describing item-passage overlap features: Correspondence between correct response and textual information (0 = No Inference, 1 = Low Level inference, and 2 = High Level Inference), Location of Relevant Information (1 = Early, 2 = Middle, 3 = Late; and 4 = Entire Passage), and Elaboration of Information (scored as the percent of text that must be processed to correctly answer the item). The resulting cognitive processing model accounted for 87% of the variance in item difficulty, with Correspondence as the strongest predictor and Elaboration an insignificant predictor.

Rupp et al. (2001) applied TBR to model listening and reading items. Despite a small sample size (84 nonnative English speakers of varying ability levels), two strengths are unique to their study. First, they employed both TBR and multiple linear regression analyses, and thus provided multiple perspectives to more fully interpret the item difficulty model. Second, when defining the predictors, they linked cognitive demands of an item to text and item features by proposing that the processing underlying task performance was associated with text features (e.g., information density), item features (e.g., lexical overlap between correct answer and distractors), and text-by-item interactions (e.g., type of match). A limitation with their study might be the lack of strong evidence for combining the items across the modalities in item modeling. They assumed that reading and listening items could be grouped according to information processing characteristics common to both modalities. A think-aloud or dimensionality analysis may help to clarify whether modeling reading and listening item groups separately would be better in terms of interpretability of the models.

Huff (2003) modeled item difficulty for the new TOEFL using TBR for the purpose of providing descriptive score reports regarding examinees' English language proficiency. In her application, the data were examinee performances on the Listening and Reading items from two prototypical parallel forms (1,372 examinees for Form 1 and 1,331 for Form 2). Her final models accounted for 56% of the variance in item difficulty for reading items and 48% for listening items. Several features distinguish her analysis. First, both dichotomously and polytomously scored items were involved. Item difficulty parameters were estimated with the 3-PL IRT model for dichotomous items and graded response model (GRM) (Samejima, 1997) for polytomous items. Second, unlike previous TBR studies where items were classified by user specifications, Huff introduced cluster and dimensionality analyses to complement the subjective judgment of item classifications. Her study showed that dimensionality analyses facilitated item grouping and substantive interpretations of item modeling solutions. Third, the predictors used in her TBR analysis were the existing item and passage codes developed by the TOEFL developers. These predictors included item and text features and were defined using Mislevy's (1994) framework of evidence-centered design and Bachman's (1990) framework of CLA. However, as these existing codes were not defined specifically for item difficulty research, factors affecting reading/listening item difficulty and the interaction between item and text—that is, what an examinee is required to do and the type of information in the text—were not taken into account.

Defining item features is the fundamental issue in applying TBR to item modeling, as the item features that are included in the model and how they are coded are closely related to model interpretability (Ewing & Huff, 2004; Huff, 2003). In the assessment of reading, assessing examinees' processes when they read passages and respond to items has been increasingly emphasized, and the methods in cognitive psychology such as task and verbal report analysis have been used to gain insights into examinees' processes during task performance (Alderson, 2000). Accordingly, identifying cognitive processes underlying reading item performance needs to consider theoretical information, cognitive features of items, and examinees' verbal reports about their item solving processes.

Method

Description of the MELAB Reading Section

According to the *MELAB Technical Manual* (English Language Institute, 2003), the reading section is designed to assess examinees' understanding of college-level reading texts. The reading section consists of four passages, with each followed by five multiple-choice items. Each item consists of a question stem and four options (one key and three distracters). Examinees are instructed to read the passages and select the single best answer based on the information in the passages. All passages are expository texts and the language characterizes English for academic purposes. The readability of the passages, as measured by a standard readability formula, suggests that the vocabulary and structural complexity of the passages are at the college level. The topics of the passages are accessible to all examinees; no prior knowledge is required to understand a passage or solve an item. To counter any possible bias towards examinees of a particular educational or cultural background, ELI-UM selects texts on a range of topics and includes different genres of passages in each test form. According to the ELI-UM item-writing guidelines, the questions following each passage are intended to assess a variety of reading abilities, including recognizing the main idea, understanding the relationships between sentences and portions of the text, drawing text-based inferences, synthesizing, understanding author's purpose or attitude, and recognizing vocabulary in context. In this study, the analyses were performed on the reading section of two parallel MELAB forms, Form E and Form F, administered during the years 2003 and 2004. The passages included in each form range from 229 to 265 words in length and are on topics in the social science, biological science, physical science, and agriculture subject areas.

Defining the Initial Cognitive Processing Model and Cognitive Variables

Following an analysis of the literature and the constructs assessed by the MELAB reading section, a theoretically supported cognitive processing model was hypothesized to underlie MELAB reading test item performance. The model was considered to have the following 10 general categories of processing components.

1. Recognize and determine the meaning of specific words or phrases using context clues or phonological/orthographic/vocabulary knowledge (PC1).
2. Understand sentence structure and sentence meaning using syntactic knowledge (PC2).
3. Understand the relationship between sentences and organization of the text using cohesion and rhetorical organization knowledge (PC3).
4. Speculate beyond the text, e.g., use background/topical knowledge (PC4).
5. Analyze the function/purpose of communication using pragmatic knowledge (PC5).
6. Identify the main idea, theme, or concept; skim the text for gist (PC6).
7. Locate the specific information requested in the question; scan the text for specific details, which includes (a) match key vocabulary items in the question to key vocabulary items in the relevant part of the text, and (b) identify or formulate a synonym or a paraphrase of the literal meaning of a word, phrase, or sentence in the relevant part of the text (PC7).
8. Draw inferences and conclusions based on information implicitly stated in the text (PC8).
9. Synthesize information presented in different sentences or parts of the text (PC9).

10. Evaluate the alternative choices to select the one that best fits the requirements of the question and the idea structure of the text (PC10).

Based on this theoretical model and empirical studies of processing difficulty for multiple-choice reading test items (Gorin, 2002; Huff, 2003; Jamieson et al., 2000; Kirsch & Mosenthal, 1990; Sheehan & Ginther, 2001), the cognitive processing features of the MELAB reading items scored for consideration in the TBR statistical model were derived. PC1, PC2, and PC3 were coded as the degree to which the corresponding process was required to solve the item (0 = Low; 1 = Middle; 2 = High). In addition, PC1 involved a variable measured as percentage of specialized and infrequent words in the part of the text where the necessary information to solve the item is located, based on the hypothesis that text with more specialized and infrequent vocabulary items will be more difficult to process, understand, and recall when answering the questions. This variable was scored using Web VP version 2.0 (Cobb, 2004). PC5 was coded on a scale from 0 to 4 with higher numbers representing more complex processing required to solve the item. PC6 and PC7 were coded on a three-point scale (0 = the question does not request locating specific details in the text; 1 = the information requested in the question can be located in the text by identifying the lexical overlap between the question and the text; 2 = the information requested in the question can be located by identifying a synonym or a paraphrase of the literal meaning of a word, phrase, or sentence in the text). PC4 and PC8 were coded as correspondence between correct response and information in text (0 = Literal or synonymous match; 1 = Low text-based inference; 2 = High text-based inference; 3 = Prior knowledge beyond text). PC9 was coded as the degree to which synthesis was required to solve the item (0 = No synthesis; 1 = Low-level synthesize; 2 = High-level synthesize). PC10 was coded as the number of distractors that contained lexical overlap with text or ideas explicitly/implicitly stated in the text. As it was hard to reach consensus on this variable, it was counted as the average of the ratings by the three raters.

Coding the MELAB Reading Items

After defining the initial model and variables, three raters coded the MELAB reading items on Form E and Form F in terms of the cognitive processes required to correctly answer each item. All raters were doctoral students in educational psychology, with experience in teaching reading to adult L2 learners. To enhance rating reliability, a rating instrument was employed, which contained components of the initial cognitive model described above, definitions of the cognitive processes covered, example cognitively based item features, and scored variables for item coding. The rating instrument also allowed the raters to indicate any processes that were not included in the rating instrument but were required for item solving.

Prior to formal rating, a group training session was held, during which the researcher introduced the study and two MELAB reading sections, acquainted the raters with the rating instrument, and clarified the rating procedure. Discussion was encouraged as a way of achieving common definitions and understanding. As part of the training, a sample passage with five associated items was provided for practice. The raters were asked to (1) answer the sample items, (2) mark their answers using the answer key, and (3) code the sample items using the rating instrument. Upon completion, the coding results and inconsistencies were discussed. Following that, the raters independently coded the reading items included in both forms. To ensure that the procedure was followed exactly, each rater was provided three envelopes, which contained instructions and materials for each step of the coding. Envelope A contained two MELAB reading sections. The raters were instructed to read the passages and answer the items as if they were taking a reading test. Upon completing this task, they were instructed to open Envelope B, which contained the answer keys provided by the ELI-UM. The raters were asked to check and correct their answers. Upon completion, they were instructed to open Envelope C, which contained the rating instrument and instructions for item coding. The raters were asked to complete the entire task in 3 days, and to return their completed work with all the instructions and materials in the original envelopes to the researcher by the end of the third day. The researcher entered the rating data collected from the raters into SPSS 13.0 (SPSS, 2005) and verified for 100% accuracy. Rater consistency was examined using generalizability theory (G-theory). G-theory offers a more comprehensive framework for studying the rater data. With G-theory, rater performance can be studied across a number of different factors, such as cognitive processes and items. Finally, a meeting was held for the raters to look at the coding summary conducted by the researcher by hand and to reach consensus on the item codes for which there was a lack of agreement. Following the meeting, the researcher entered the consensus codes into the Microsoft Excel 2000 and verified to ensure 100% accuracy.

Validating the Cognitive Processes through Verbal Reports

To ensure that the cognitive model and the associated cognitive variables are faithful descriptors of examinees' cognitive processes, concurrent and retrospective verbal reports were collected from individual participants as they worked through the MELAB reading items. The participants were 10 Chinese-speaking students (4 male, 6 female) enrolled in an undergraduate or graduate program at the University of Alberta in fall 2005. They ranged in age from 19 to 32, received at least 11 years of basic education and at least eight years of English education in China, and had resided in English-speaking countries for no longer than six months. The participants were randomly assigned to take Form E and Form F, with an equal number of participants for each form.

Data collection took the form of administering Form E or Form F of the MELAB reading test and asking participants to report, in Chinese, English, or both, what they were thinking as they answered the items and what they thought while answering the items after completing each item. To avoid the possibility that researcher probes could lead the participants, nonmediated verbal reports were used. Given that the original form containing 20 items was too long for both concurrent and retrospective verbal reports, data from each participant were collected during two separate sessions scheduled on 2 different days within a week, with the first 10 items administered on day one and the second 10 items on day two.

On day one, the researcher met with a single participant in an empty office at the university. The researcher and participant sat side by side at a table on which there were a digital audio recorder, a microphone, and a folder containing the experimental materials. These materials included a sheet of directions, two practice questions, and the test form. The researcher first explained the nature and procedure of the task. Given that participants may not have been familiar with the verbal report methods, the researcher provided an opportunity for them to practice verbal reporting skills, using two questions presented in Ericsson and Simon (1993). The researcher asked them to report aloud their thinking and the information they were attending to while answering the sample items (concurrent reports). After selecting an answer to an item, the researcher asked them to report their remembrance about their thoughts and the places they were attending to from the time they began to read the question until they selected an answer (retrospective reports). The participants were asked to answer the items as if in a real testing situation and to verbalize whatever was on their minds while and after completing each item. The participants practiced the verbal report procedures using the sample questions. If the participants remained silent for a lengthy time period, they were reminded to keep talking. Once the participants became accustomed to the reporting procedures and had no more questions, they were administered the first two passages with their associated 10 items from Form E or Form F and the digital audio recorder was turned on. Participants were asked to read the passages silently, verbally express their thought processes while responding to the items, and upon completing each item, retrospectively describe aloud what they remembered about their thought processes used to answer the item. If the participants remain silent for a lengthy time period, they were prompted to keep talking. On day two, following the same procedure, the participants completed the remaining two passages with their accompanying 10 items. To maximize the consistency among the sessions, standardized procedures and instructions were followed for each session.

The participants' verbal reports were transcribed verbatim and typed into the computer for analysis. The researcher reviewed all verbal reports and coded them for the cognitive processes used to answer each item. The initial cognitive processing model was used as a starting point for classifying the verbal report data. Statements or phrases in the reports associated with each cognitive process were segmented and assigned a code. Additional processes gleaned from the transcripts were categorized and added to the model. After the verbal report data were coded and the additional processing categories developed, the cognitive processing model was revised as necessary and then used as a coding scheme to recode the previously coded data by the researcher. To evaluate the coding reliability, an independent rater (i.e., a colleague of the researcher, who has comparable expertise as the researcher and no experience with the study) was invited to code 40% of the verbal report data. The independent rater was first trained in the data coding procedures. During the training, the researcher discussed the coding scheme (i.e., the list of 10 processing components in the initial cognitive model with definitions and examples of each) with the rater, demonstrated the coding, and provided a chance to practice using the verbal reports from one of the participants. After the training, the rater independently coded four randomly selected verbal reports, using the coding scheme, and then his codes were compared to those of the researcher. The percent of interrater agreement was calculated to evaluate the consistency of the verbal report coding.

To determine the final set of item features, the cognitive processes obtained through the analysis of verbal reports were compared to those obtained through rater coding. Consistent findings were checked, complementary findings combined, and contradictory findings replaced with the cognitive processes inferred from the verbal reports. Based on the results of the comparison, the cognitive model and the rating instrument were refined as necessary. Next, the three previous raters met together with the researcher to review the changes about the model and the rating instrument, and to recode the items using the modified rating instrument. Once consensus was made, the final list of the cognitive processes required to correctly answer each item was formatted into two 20 x *k* matrices (20 is the total number of items on each test form and *k* is the number of cognitive processes required to correctly answer the items), with one matrix for Form E and the other for Form F.

Developing Item Difficulty Model

Data

Two data files containing examinee item responses to reading items on Form E and Form F of the MELAB were provided by the ELI-UM. One data file contained item responses from 1703 examinees on Form E administered from January 2003 through September 2004. The other data file contained item responses from 1044 examinees on Form F administered from January 2003 through October 2004. Neither file contained missing data, since the examinees who did not attempt one or more of the items (3.2% of the total number of examinees) had been excluded, given the consideration that these examinees may have been simply guessing and thus were not instigating the processes required by item solution (J. Johnson, personal communication, January 18, 2005).

Date Scoring and Analysis of Psychometric Characteristics

Examinee response data were exported into SPSS 13.0 (SPSS, Inc., 2005). Items were scored to the key, with 0 representing the incorrect responses and 1 representing the correct responses. Descriptive statistics and reliability estimates were computed for the two reading sections using the computer program Lertap 5 (Nelson, 2000). Given the lack of local item independence due to common passages (Kolen & Brennan, 2004), item parameter estimates were calibrated using the testlet response theory (TRT) model (Wang, Bradlow, & Wainer, 2002). The TRT is a four-parameter dichotomous IRT model that introduces a testlet effect parameter, $\gamma_{ig(j)}$. The TRT model is expressed as:

$$p(y_{ij} = 1 \mid \theta_i) = c_j + (1 - c_j)\frac{\exp[a_j(\theta_i - b_j - \gamma_{ig(j)})]}{1 + \exp[a_j(\theta_i - b_j - \gamma_{ig(j)})]},$$

where y_{ij} is the score for examinee *i* on item *j*, θ_i is the ability level of examinee *i*, $p(y_{ij}= 1|\theta_i)$ is the probability that examinee *i* at the ability level θ correctly answers item *j*, a_j is the discrimination parameter of item *j*, b_j is the difficulty parameter of item *j*, c_j is the pseudoguessing parameter of item *j*, and $\gamma_{ig(j)}$ is the testlet effect parameter indicating the testlet effect for examinee *i* responding to item *j* that is nested in testlet *g*. The TRT model separates the testlet effect from examinee ability by estimating the testlet effect parameter (γ) for each testlet and each examinee during the calibration of the *a*-, *b*-, and *c*-parameters. In this way, the problem of local dependence in passage-based reading tests is attended to and the resulting item parameter estimates are more accurate (Wang et al., 2002). For this study, the parameters of the reading items were estimated separately within Form E and Form F, using the computer program SCORIGHT 3.0 (Wang, Bradlow, & Wainer, 2004). This program is used because it is based on the TRT model and can address the problem of local dependence. The item difficulty parameter estimates calibrated within each form were formatted into two 20 x 1 vectors, with one vector for Form E and the other for Form F.

TBR analysis

To determine the extent to which the identified cognitive processes accounted for the item difficulty estimates, two sets of TBR analyses were performed using the regression tree module available through SPSS 13.0. One set of TBR analysis was performed on Form E as the principal analysis, and the other performed on Form F to cross-validate the results. In both sets of the TBR analyses, the predictors were the 20 x *k* cognitive processes matrix, and the criterion the 20 x 1 vector of the item difficulty estimates for the corresponding form. The analysis of Form E began with the placement of the 20 items in a single node at the top of the tree, where 0% of the variance was explained. The items were successively split into increasingly homogeneous clusters, according to the classification of the cognitive processes required for item solution. At the bottom of the tree, each item was classified into its own cluster, where 100% of the difficulty variance was explained. At each stage of splitting, a recursive partitioning algorithm (Breiman, Friedman, Olshen, & Stone, 1984) was used to evaluate all possible splits of the predictor variables. The best split was the one that resulted in the largest reduction in the deviance between the parent node and the sum of the two child nodes. The smaller the deviance value is, the more homogeneous the items within a node are. Generally, increasing the level and terminal nodes of the tree would lead to an increase of the explained variance in the difficulty. However, in order to determine the levels of the tree and the number of terminal nodes the parsimony and interpretability of the model needs to be taken into account. If adding a new level and more terminal nodes does not contribute to the improvement of the variance explained, the more parsimonious model is preferred (Huff, 2003). Hence, the final stage of the TBR analysis was to increase the parsimony and interpretability of the model by pruning, which involved removing one or more sets of child nodes and collapsing the similar terminal nodes. After completing the principal analysis with Form E, cross-validation was performed using the 20 items on Form F through the same procedure. If the final item difficulty model obtained using the Form E data can be replicated using the Form F data, then the TBR analyses will provide strong empirical support for the cognitive processing model underlying the MELAB reading test items.

Results

Coding the MELAB Reading Items

Rater consistency was examined using a G-study fully crossed item by skill by rater mixed effect design, in which items were treated as the object of measurement, raters a random facet, and cognitive processes a fixed facet. The computer program GENOVA (Crick & Brennan, 1983) was used to obtain the variance components and reliability coefficient, which were displayed in Table 1.

Table 1. Variance Components and Reliability Coefficient from the G-Study

Effects	Degrees of Freedom	Variance Components	Percent
Item	39	0.1144	11.50
Rater	2	0.0455	4.57
Process	7	0.0643	6.46
Item-Rater Interaction	78	0.0703	7.07
Item-Process Interaction	273	0.3324	33.42
Rater-Process Interaction	14	0.0401	4.03
Residual	546	0.3276	32.94
Reliability			0.75

Several notable things can be observed from Table 1. First, the reliability coefficient was 0.75, which indicates that the items were consistently coded by the raters. For a more comprehensive understanding of the raters' performance, the different effects involving raters can be referred to. Among all effects but residual, the effects involving raters accounted for a negligible amount of the total variance. Only 4.57% of the total variance was accounted for by the rater effect, 7.07% by the item-rater interaction, and 4.03% by the rater-process interaction. Hence, it can be concluded that the raters performed consistently across processes and across items. Second, the largest variance component came from the item-process interaction, indicating that different processes were required to solve different items. Third, item effect only accounted for 11.50% of the total variance, which indicates that the average ratings received by the items across different processes were comparable, and that an item receiving a high rating on one process might receive low ratings on other processes.

Results of the item coding showed several major points. First, all components of the initial cognitive processing model were involved in solving the reading items on Form E and Form F, and no additional processes were raised by the raters. Second, correctly answering an item requires multiple cognitive processes. Third, correctly answering an item was often associated with text/item features, knowledge of particular lexical items, drawing inferences, and evaluating alternative options. The final set of consensus codes is presented in Appendix A. To compare the cognitive item features across forms, descriptive statistics for the consensus codes were calculated and are presented in Table 2. As the table shows, the distributions of item features were comparable across forms. For the features "applying pragmatic knowledge" and "locating information," the mean ratings for the items on both forms were identical. For the features "percentage of specialized and infrequent vocabulary in relevant part of the text" and "the number of plausible distractors," the mean ratings for Form F items were slightly lower than those for Form E items, but for the remaining five features, the mean ratings for Form F items were slightly higher than those for Form E items.

Table 2. Descriptive Statistics for Consensus Item Codes

Process	Word Recog.	% Sp. Words	Syx	Text Org.	Prag-matic	Locate	Infer.	Syn-thesis	Dis-tractor
Form E									
N	20.00	20.00	20.00	20.00	20.00	20.00	20.00	20.00	20.00
Min.	0.00	10.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Max.	2.00	41.67	2.00	2.00	4.00	2.00	3.00	2.00	3.00
Mean	1.30	22.04	1.45	0.80	2.00	1.20	1.20	0.70	1.90
SD	0.73	7.62	0.69	0.95	1.56	0.70	0.77	0.80	0.91
Form F									
N	20.00	20.00	20.00	20.00	20.00	20.00	20.00	20.00	20.00
Min.	1.00	8.70	1.00	0.00	0.00	0.00	0.00	0.00	0.00
Max.	2.00	31.58	2.00	2.00	4.00	2.00	2.00	2.00	3.00
Mean	1.50	19.86	1.60	1.05	2.00	1.20	1.25	1.25	1.65
SD	0.51	5.98	0.50	0.76	1.41	0.70	0.79	0.72	1.14

Verbal Reports of the Cognitive Processes

The reliability of assigning processes to the various processing categories was evaluated. An independent rater coded four verbal reports, two of which were randomly selected from the Form E participants and two randomly selected from the Form F participants. The coding results of the independent rater were compared to those of the researcher. Consistency was defined as the extent to which the verbal report data segments were coded using the same processing categories by both raters. Of a total of 291 processes coded by both raters, 247 agreements occurred. Hence, the percentage of total agreement between the researcher and the independent rater was 85%. The total number of the cognitive processes codes assigned to each verbal report is presented in Table 3.

Table 3. Cognitive Processes Frequencies for Each Participant and Form

Process /Partici.	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9	PC10	Other	Total
E1	9	6	6	1	5	3	18	8	3	12	1	72
E2	4	4	5	0	1	2	17	10	2	11	2	58
E3	2	2	3	0	1	2	15	5	4	15	4	53
E4	6	1	0	1	0	3	12	3	1	3	2	32
E5	1	3	4	0	0	2	15	9	3	14	1	52
Total	22	16	18	2	7	12	77	35	13	55	10	267
F1	7	4	1	2	0	6	15	11	7	13	2	68
F2	6	5	0	2	3	4	9	4	5	8	2	48
F3	3	4	2	0	3	4	16	8	2	13	2	57
F4	4	9	6	0	1	7	15	12	3	13	1	71
F5	6	3	1	1	3	5	12	8	5	9	4	57
Total	26	25	10	5	10	26	67	43	22	56	11	301

PC = Processing Component; component in the initial cognitive processing model.

Table 3 provides insights into the cognitive processes used by the participants while they were answering the MELAB reading items on Form E and Form F. For both forms, the cognitive process most frequently inferred from the verbal reports was PC7 (scanning for details/matching the question to the relevant information in the text). The participants taking Form E reported this process 77 times in total and the participants taking Form F reported 67 times in total. The second and the third most frequently reported processes were PC10 (Evaluate alternative options) and PC8 (drawing text-based inference), respectively. More cognitive processes can be inferred from the Form F participants' verbal reports than from the Form E participants' verbal reports.

Among the participants taking Form E, E1 reported the highest number of processes (a total of 72), and these processes covered all categories in the proposed cognitive processing model. The processes most frequently reported by this participant included PC7 (scanning details), PC10 (evaluating alternative options), PC1 (identifying word and word meaning), and PC8 (drawing inferences). This participant correctly answered 16 of the 20 items. Among the participants taking Form F, F4 reported the highest number of processes (a total of 71) and these processes covered all categories in the cognitive model but PC4 (speculating beyond the text). The processes frequently reported by this participant included PC7 (scanning details), PC10 (evaluating alternative options), PC8 (drawing inferences), and PC2 (using syntax knowledge). This participant correctly answered 19 of the 20 items. The participants reporting the lowest number of processes on each form were the ones who scored the lowest in the group of participants taking that form. Participant E4 reported a total of 32 processes and scored 13 correct out of 20 on the Form E reading section, and participant F2 reported a total of 48 processes and scored 11 correct out of 20 on the Form F reading section.

Additional processes obtained from the participants' verbal reports can be classified into three categories. The first category includes metacognitive and metalinguistic strategies, such as *deciding an answer after all options are evaluated*, *translating into Chinese, going back and forth between text and items*, *marking the text as reading to help locate the information when answering the questions, skipping specialized nouns*, *answering easier items first*, *being aware of the processes used*, *analyzing what the question assesses,* and *switching to other processes (e.g., evaluating options) to save time when one process doesn't work (e.g., the required information can't be found in text).* The second category includes construct-irrelevant processes, such as *random guessing* or *guessing based on a constructed situation model or prior knowledge*. The third category is related to affect and memory, such as "I find it hard to concentrate at the beginning," "I like scientific text," and "I cannot remember where I read this in the text." While these data provided invaluable insights into examinees' item solving processes, they were not added to the initial cognitive processing model, given the considerations that (1) the use of these processes in item solving varied from person to person and from item to item, (2) they were hard to code for consideration in statistical models, and (3) they were not included in the constructs assessed by the MELAB reading section.

Next, to validate the components of the cognitive model and the item features coded by the raters, the processes used by the participants who correctly answered each item were summarized and compared to the final set of consensus codes obtained from item coding. The results of this comparison are presented in Appendix C. An examination of Appendix C reveals several major points. First, the cognitive processes inferred from the participants' verbal reports provide evidence that correctly answering an item often requires multiple processes. Second, the processes used to correctly answer the reading items on both forms covered all components of the proposed cognitive processing model. Third, the processes reported by the participants who correctly answered each item supports the coding of the MELAB items in terms of the cognitive processes required to correctly answer each item. Of a total of 160 features coded for the reading items on Form E (20 items x 8 variables), 117 item features (73.1%) are supported by the verbal report data. When the process of using pragmatic knowledge, which was reported infrequently by the participants, is excluded, 112 of the remaining 140 item features (80.0%) are supported by the verbal data. Likewise, of a total of 160 features coded for the reading items on Form F (20 items x 8 variables), 111 item features (69.4%) are supported by the verbal data, and when the process of using pragmatic knowledge is excluded, 106 of the remaining 140 item features (75.7%) are supported by the verbal report data. Hence, it is considered that the item features were reasonably coded and no further modifications were made to the final set of consensus item codes.

For each cognitive variable coded by the raters, the total number and proportion of items for which the verbal data included that feature are presented in Table 4. Proportion is defined as the number of items for which the coding could be validated by the verbal data divided by a total of 40 items coded on that variable. As Table 4 shows, the highest degree of correspondence between the item coding and the verbal data occurs on the variable Locate. All 40 items on both forms coded for this feature are supported by the verbal data. The lowest degree of correspondence between item coding and verbal data occurs on the variable Pragmatic. Of a total of 40 items coded for this feature, only 10 items could be validated by the verbal data. Generally, there is more overlap between the item codes and verbal data for the variables Locate, Distractor, Inference, and Synthesis than that for the variables Word, Syntax, Text Organization, and Pragmatic.

Table 4. Correspondence between Item Codes and Verbal Report Data

Process	Word Recog.	Syntax	Text Org.	Prag-matic	Locate	Infer.	Syn-thesis	Dis-tractor
f	27	21	27	10	40	34	33	35
%	67.5	52.5	67.5	25.0	100.0	85.0	82.5	87.5

Results of Item Difficulty Modeling

Psychometric Characteristics

The psychometric characteristics of the MELAB reading items on Form E and Form F are summarized in Table 5. As the table shows, the psychometric characteristics of the two sections are comparable, though the reliability of the reading section on Form F is slightly lower than that of the reading section on Form E. The empirical data supports the parallelism of the two sections. Item parameter estimates were calibrated using the testlet response theory (TRT) model (Wang et al., 2002). Item difficulty parameter estimates for the reading items on Form E and Form F are presented in Appendix B.

Table 5. Descriptive Statistics and Reliability for the Two Reading Sections

Form	N_{scores}	Minimum	Maximum	Median	Mean	SD	Reliability
E	1703	0.00	20.00	11.00	10.94	4.19	0.79
F	1044	1.00	20.00	11.00	10.71	3.65	0.71

TBR Analyses

Separate TBR analyses were conducted on Form E and Form F. Both analyses started with nine predictors.

Form E. For Form E, five of the nine predictors entered the tree: Distractor (number of plausible distractors), Pragmatic (pragmatic knowledge), Syntax (syntax knowledge), Text Org. (knowledge of text organization), and Speword (proportion of specialized and infrequent words in the part of the text where the information for answering the question is located) (See Figure 1). Taken together, these five variables accounted for 90.4% of the total variance in item difficulty.

As some important predictors may be masked in the tree-building process, it is crucial to inspect the importance of the predictors to the model (Breiman, et al, 1984). Table 6 presents the importance of individual predictors in the item difficulty model built for Form E. As the table shows, the most important predictor in the model is Distractor. However, the predictor Inference, which did not appear in the model, is the second most important predictor, and is far more important than the remaining predictors. It is highly likely that this predictor was masked in the tree-building process and given the importance of the predictor Inference, it appeared unwarranted to exclude it from the model. Consequently, a new tree-building process was undertaken by successively adding in predictors based on their importance to the model. According to the statistical principle of parsimony (e.g., Kerlinger, 1979), two stopping rules were used: (1) the newly added predictor did not lead to a significant increase in the explained variance of item difficulty, and (2) the total variance in item difficulty was maximally explained. The model that explained the largest amount of variance in item difficulty with the least number of predictors was considered the most parsimonious and therefore used for interpretation. First, the most important predictor, Distractor, was used in the model and this predictor explained 41.4% of the total variance in item difficulty. Next, the predictor Inference was entered into the model. However, this predictor did not lead to any increase in the explained variance and 60% of the variance in item difficulty was left unexplained. It was likely that Inference was again masked in this tree-building process. Subsequently, the next three important predictors, Pragmatic, Speword, and Synthesis, were successively fed into the model. Table 7 displays the contribution of the top five important predictors to the explained variance in item difficulty.

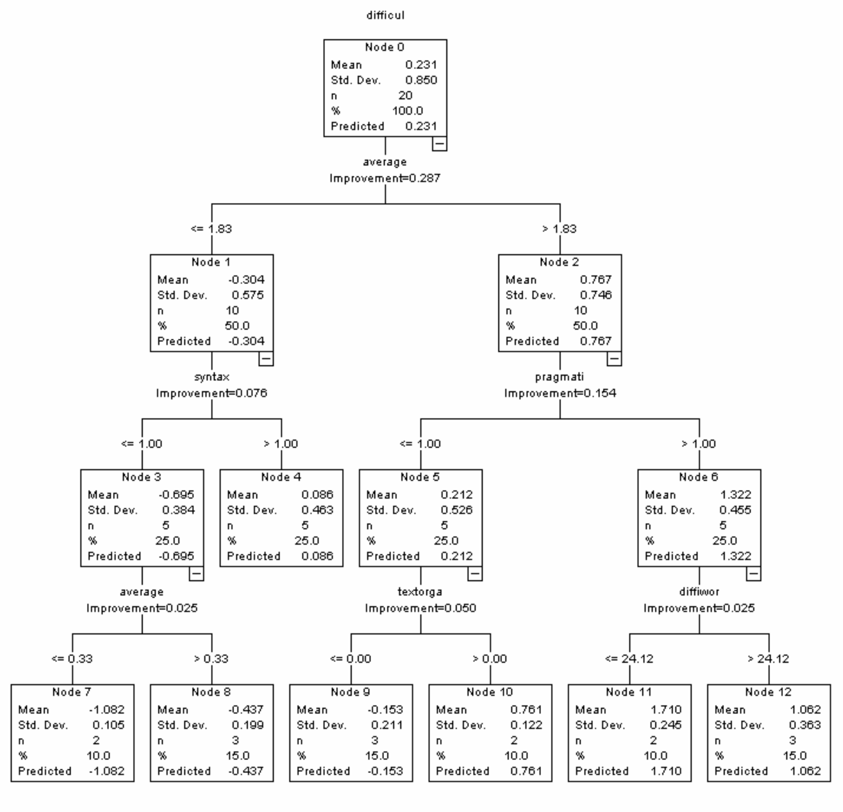

Figure 1. Initial Tree Diagram for the Reading Items Included in Form E.

Table 6. Importance of Individual Predictors in the Item Difficulty Model for Form E

Predictors	Importance	Normalized Importance (%)
Distractor	0.422	100.0
Inference	0.329	77.9
Pragmatic	0.207	49.1
Speword	0.169	40.0
Synthesis	0.122	28.9
Syntax	0.120	28.5
Locate	0.100	23.7
Text Org.	0.096	22.8
Word Recog.	0.049	11.7

Table 7. The Contribution of the Predictors to Explained Variance on Form E

Predictors	Total Variance Explained (%)	Unique Variance Explained (%)
Distractor	41.4	41.4
Inference	41.4	0.0
Pragmatic	85.8	44.4
Speword	90.7	4.9
Synthesis	91.1	0.4

As Table 7 shows, the predictor Inference did not increase the explained variance in item difficulty. However, when the predictors Inference and Pragmatic were fed into the model, a drastic increase was achieved in the explained variance (44.4%). Given the importance of the predictors Inference and Pragmatic and the contribution of both predictors to the improvement of the model, this increase in the variance explained was likely from the joint contribution of Inference and Pragmatic. The predictor Speword explained an additional 4.9% of the total variance in item difficulty. However, when the predictor Synthesis was fed into the model, there was virtually no increase in the explained variance (0.4%). Therefore, for Form E, the tree was built with four predictors: Distractor, Inference, Pragmatic, and Speword, which accounted for 90.7% of the total variance in item difficulty. Figure 2 presents the tree diagram for Form E with four predictors.

Figure 2 presents the mean, standard deviation, and the number of items for each node. As the figure shows, Distractor produced one split at the first level and one at the second level. Both splits indicated that the items with more plausible distractors tended to be more difficult than the items with less plausible distractors. At the second level, Pragmatic produced a split, indicating that the items requiring more pragmatic knowledge (e.g., analyzing author’s opinion and extrapolation) tended to be more difficult than the items requiring less of such knowledge (e.g., facts). Similarly, at the third level, Inference produced a split, indicating that the items requiring high text-based inference or speculation beyond text tended to be more difficult than the items requiring no or low text-based inference. At the third level, Speword produced another split, indicating that the items requiring processing the part of the text containing more specialized and infrequent words tended to be less difficult than the items requiring processing the part of the text containing less of such words.

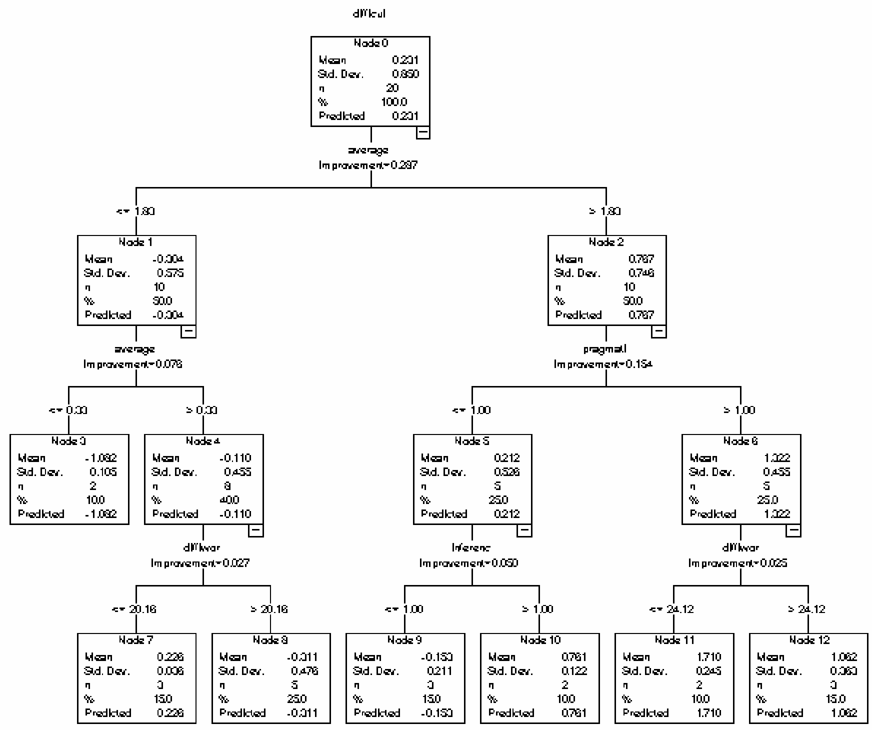

Figure 2. Tree Diagram for the Form E Reading Items.

Form F. For Form F, three of the nine predictors entered the tree: Distractor, Inference, and Syntax. Taken together, these three predictors accounted for 94.5% of the total variance in item difficulty. Figure 3 presents the mean, standard deviation, and the number of items for each node. As the figure shows, Distractor produced the first split, separating the 20 items into two groups with different mean item difficulties. This split again indicated that the items with more plausible distractors tended to be more difficult than the items with less plausible distractors. At the second level, a split was made based on the predictor Inference, which again indicated that the items requiring high text-based inference tended to be more difficult than the items requiring no or low text-based inference. The predictor Syntax produced one split at both the second and the third level. These two splits indicated that the items requiring knowledge of complex or infrequent sentence structure tended to be more difficult than the items requiring knowledge of simple sentence structure.

The importance of the predictors to the model for Form F was inspected and is presented in Table 8. As the table shows, the three predictors entering the tree model (i.e., Distractor, Inference, and Syntax) were the three most important predictors to the model. Hence, the tree built with the three predictors was taken as the final model for Form F.

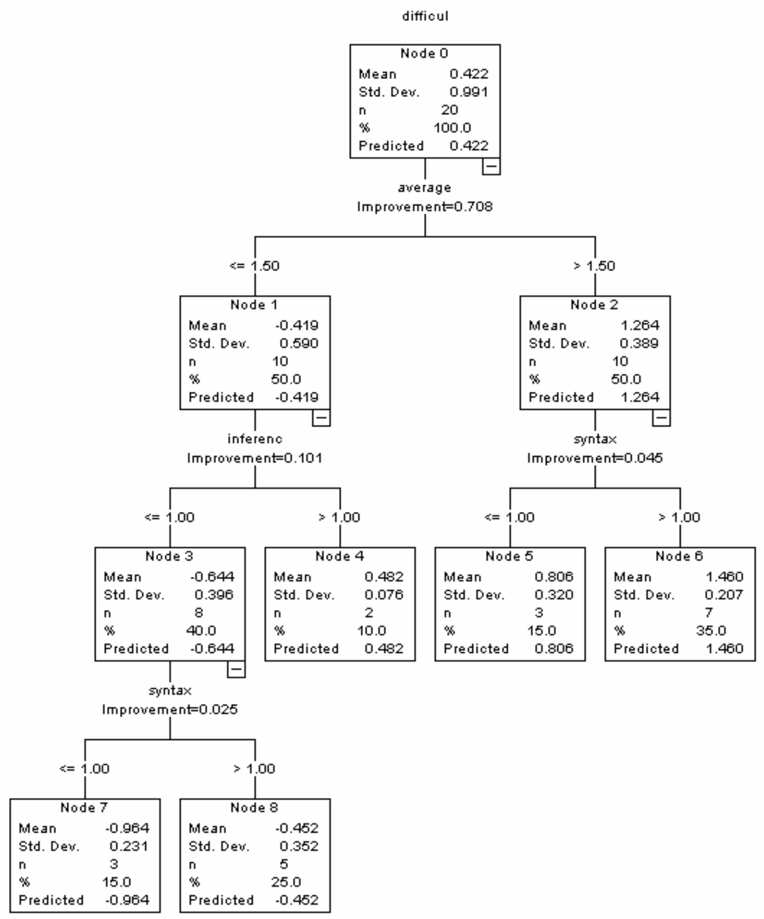

Figure 3. Tree Diagram for the Form F Reading Items.

Table 8. Importance of Individual Predictors to the Model for Form F

Predictors	Importance	Normalized Importance (%)
Distractor	0.792	100.0
Inference	0.444	56.0
Syntax	0.157	19.8
Pragmatic	0.145	18.3
Speword	0.135	17.1
Locate	0.082	10.3
Word Recog.	0.040	5.1
Synthesis	0.012	1.5
Text Org.	0.007	0.9

Discussion

In this study, a three-pronged procedure was employed to develop and test a cognitive processing model hypothesized to underlie MELAB reading item performance. First, theoretical information regarding the L2 reading processes and reading ability constructs were reviewed. Next, to provide clear, faithful, and informative definitions of the cognitive processes involved in solving the MELAB reading items, cognitive demands of the items were analyzed and the cognitive processes that examinees might use to correctly answer the items were investigated. Finally, the proposed cognitive processes were validated through empirical evaluation of objective performance on the MELAB reading items using a cognitively based measurement model called tree-based regression.

Summary and Discussion of the Findings

Research question 1: What cognitive processes are required to correctly answer the MELAB reading items?

Three raters independently coded the MELAB reading items on Form E and Form F in terms of the cognitive processes required to correctly answer each item. An examination of the rater consistency using G-theory indicated a fairly high level of rater agreement ($\rho = 0.75$), given that only three raters were used. Contrary to the results in Alderson and Lukmani (1989) and Alderson (1990), this finding appears to support that raters can reach agreement on the cognitive demands of an item. It appears that the use of rater training, a clearly defined rating instrument, extensive discussion, and exemplification of item coding in this study contributed to the agreement among the raters. Results of item coding show that correctly answering an item requires multiple cognitive processes, which provides evidence that solving a reading item involves simultaneous use of different cognitive components (Gorin, 2002). Results of item coding show that the cognitive processes required to correctly answer the MELAB reading items include word recognition skills, knowledge of syntax and text organization, pragmatic knowledge, skimming the text for gist, scanning the text for specific details, drawing inferences, synthesis, and evaluating alternative options, and that different cognitive processes are involved in solving different items. Results of the item coding in terms of the cognitive processes required to correctly answer the MELAB reading items are consistent with the constructs assessed by the MELAB reading section and support the cognitive processing model proposed in this study.

Research question 2: What cognitive processes are actually used by examinees when they correctly answer the MELAB reading items? How are they related to the findings in response to question 1?

Results of the verbal report data analysis show that the cognitive processes that examinees might use to correctly answer the MELAB reading items include the use of word recognition skills, knowledge of sentence and text structure, prior knowledge, pragmatic knowledge, skimming the text for gist, scanning the text for specific details, drawing inferences, synthesis, evaluating alternative options, metalinguistic and metacognitive strategies, and testwiseness. For both MELAB forms, using prior knowledge beyond text and using pragmatic knowledge were the two cognitive processes least frequently reported by the participants. Given that using prior knowledge is irrelevant to the construct assessed by the MELAB reading items, it is no wonder that this process was reported least frequently.

A comparison of the cognitive processes coded for each item to the cognitive processes inferred from the verbal reports for each item found a high degree of match between the two sources of data. Of a total of 320 item features coded for both forms (40 items x 8 cognitive variables), 228 item codes (71.3%) matched the cognitive processes inferred from the verbal reports for the corresponding item. The match occurs more frequently on the processes of locating/scanning specific details, evaluating alternative options, inference, and synthesis than on the processes of identifying word meaning, using text organization, syntactic, and pragmatic knowledge. The inconsistencies between the cognitive processes coded by the raters and the cognitive processes inferred from the verbal reports should not lead to hasty judgments about the untrustworthiness of the raters' coding. Leighton (2004) warned that verbal reports were sensitive to the demands of the task, and that they were difficult to obtain when "the task used to elicit the reports was exceedingly difficult or called upon automatic processes" (p. 12). The participants in this study were advanced-level adult L2 learners who need to use English for university-level academic studies. They were considered to (1) have mastered the basic word recognition skills, vocabulary, sentence structure, text organization, and pragmatic knowledge required for reading academic text in English, and (2) be literate in their L1 and able to use various cognitive strategies already developed from reading in their L1 to facilitate their reading in L2 (Koda, 2005; Urquhart & Weir, 1998). Hence, it is likely that the processes related to basic English language knowledge, such as word recognition skills, sentence and text structure knowledge, and pragmatic knowledge, have become automatic to this group of participants, while the processes related to cognitive skill and problem-solving strategies, such as locating specific information, evaluating alternative options, inference, and synthesis, were consciously used by the participants when answering the MELAB reading items. Given that the controlled processes rather than the automatic processes are accessible for description through verbal reporting, analyzing the cognitive demands of an item before collecting the verbal reports anticipated the automatic and the controlled processes evoked by the test items and provided valuable information to supplement the verbal report data (Ericsson & Simon, 1993; Leighton, 2004). A combination of cognitive analysis of the items and verbal reports in the current study provided an opportunity to triangulate the processes involved in item solving, and to better determine the components of the cognitive processing model and the item features for consideration in the statistical model.

Research question 3: To what extent do the cognitive processes used to correctly answer the MELAB reading items explain the empirical indicators of item difficulty?

The TBR analysis on the two forms did not converge. For Form E, four predictors explained 90.7% of the total variance in item difficulty. These four predictors were Distractor, Inference, Pragmatic, and Speword, which were, respectively, related to the cognitive processes of evaluating alternative options, drawing inferences, using pragmatic knowledge, and processing academic text with specialized and infrequent words. The results of the TBR analysis on Form E indicated that the items requiring higher level reasoning skills to make decisions regarding the response options, advanced pragmatic knowledge, and processing texts with fewer specialized and infrequent words tended to be more difficult. The finding about specialized and infrequent words in text appears counterintuitive and needs to be replicated using other test forms. The verbal report data may shed some light on the reason for this finding. The participants E1, E4, E5, and F3 all indicated that specialized and infrequent words, especially nouns, did not affect their reading or item solving, as such words could be skipped during reading and used as key words to locate the requested information in the text when answering the items.

For Form F, three predictors explained 94.5% of the total variance in item difficulty. These three predictors were Distractor, Inference, and Syntax, which were, respectively, related to the cognitive processes of evaluating alternative options, drawing inferences, and using syntax knowledge. The results of the TBR analysis on Form F indicated that items requiring higher level reasoning skills to make decisions regarding the response options and knowledge of complex sentence structures tended to be more difficult.

The inconsistent prediction of item difficulty across the forms indicates that item features likely differed among test forms due to the passages used and the nature of items included. This finding speaks for the complexity of item analysis and reminds us that caution needs to be exerted when interpreting reading performance on different test forms. Results of this study showed that while the statistical properties (e.g., descriptive and reliability) of Form E and Form F supported their parallelism, the cognitive processes elicited by the items on the two forms were not identical. Hence, besides analyzing the statistical properties of a test, substantive evidence regarding the nature of constructs assessed by the test needs to be sought to better understand the validity of the test. In addition, to ensure parallelism of test forms, tests may be constructed based on predetermined cognitive processes defined from a cognitive model. As Gorin (2002) recommended, an effective strategy for constructing and evaluating reading test items may be integrating statistical analysis with substantive analysis of the items.

While the two TBR analyses conducted in this study produced somewhat divergent results, both TBR models were relevant to the theoretical constructs of the MELAB reading section and accounted for a substantial amount of the variance in item difficulty. In addition, the pattern of agreement between the two analyses shed some light on which of the construct-relevant item features most likely affected the performance on the MELAB reading items, which could be used to guide test development and item analysis. For example, both TBR analyses indicated that the items with more plausible distractors tended to be more difficult than the items with less plausible distractors, and that the items requiring high text-based inference or speculation beyond the text tended to be more difficult than the items requiring no or low text-based inference. Such item features were consistent with the components of evaluating alternative options to decide the one that best fit, drawing text-based inferences, and speculating beyond the text in the cognitive processing model proposed in this study. In this sense, the TBR models provided evidence that this cognitive processing model was capable of describing the cognitive processes underlying the MELAB reading item performance and suggesting cognitively based mechanisms for designing new reading items (Gorin, 2002; Embretson, 1999).

Practical Implications

The cognitive processing model proposed in this study has implications for the construct validity of MELAB reading as a measure of L2 reading proficiency required for college-level academic study (Embretson, 1998; Gorin, 2002; Huff, 2003). In addition, results of this study may guide test developers to design cognitively based reading items (Enright, Morley, & Sheehan, 2002). As Gitomer and Rock (1993) suggest, “improved test design consists of building items that are constructed on the basis of an underlying theory of problem-solving performance” (p. 265). Most importantly, results of this study can be used to develop descriptive score reports and lay a foundation for the MELAB as a diagnostic measure. The TBR item difficulty model developed in this study produced clusters of items requiring similar cognitive processes. By summarizing examinee performance against item clusters, the TBR item difficulty models can be used to generate group- and examinee-level proficiency profiles (Sheehan, 1997). In this manner, large-scale language testing programs will be able to provide more meaningful feedback to score users about examinees' strengths and weaknesses in particular reading skills, suggest areas for improvement, and target instruction to individual needs (DiBello & Crone, 2001; Huff, 2003; Sheehan, 1997; Wainer, Sheehan, & Wang, 2000).

Limitations and Directions for Future Research

One limitation regarding this study is that only 40 reading items included in two forms were used to develop the cognitive processing model. To obtain reliable item features affecting MELAB reading item performance, a larger number of items from more test forms need to be examined. A second limitation is that the item features were coded by three raters having experience in teaching reading to adult ESL/EFL learners and validated using a small group of Chinese students enrolled in a university-level program. Hence, the item features obtained may not represent the cognitive processes of examinees from other language backgrounds and proficiency levels. As item difficulty is affected by the interaction between examinee and test task (Bachman, 2002), it is highly likely that item difficulty varies across language groups. Therefore, a promising area of research for the MELAB is to examine item difficulty conditioned on language background to determine whether the items perform differently for different language groups. In addition, a larger sample size for testing the cognitive processes through verbal reports may reveal more meaningful information and increase the correspondence between the item features coded by raters and the item features inferred from verbal reports. A third limitation is that the cognitive model proposed in this study did not include metacogntive and metaliguistic strategies reported by the verbal report participants, given the difficulty in coding and including such item features in statistical models. Future research may examine the relationship between such strategies and reading item difficulty, and include them in the cognitive processing model. Finally, this study validated the proposed construct by examining empirical indicators of item difficulty. Modeling other item statistics, such as item discrimination, in terms of the cognitive processes involved in item solving may reveal more meaningful information for test developers.

Conclusion

In this study, a model of cognitive processes underlying MELAB reading item performance was developed and tested. The model linked substantive theories in the domain of L2 reading to the MELAB reading items. The embracement of theoretical information regarding L2 reading processes and substantive analysis of the reading items, which is lacking in current research on the MELAB, will make possible theory-based test development and score interpretations. Moreover, the integration of cognitive theories on L2 reading and a cognitively based measurement model contributes to our understanding of the relationship between item features and item difficulty, informs the design of cognitively based reading items, and lays a foundation for the MELAB as a diagnostic measure. Finally, the three-pronged procedure used to develop and validate the cognitive model, that is, analysis of an item's cognitive demands to explore the automatic versus controlled processes evoked by test items, collection of verbal reports to investigate the actual cognitive processes used by examinees when answering test items, and TBR to model item performance, promotes the union of cognitive psychology and assessment in the field of second/foreign language testing.

Acknowledgements

I would like to express my sincere thanks to the English Language Institute of the University of Michigan for funding this project, providing the data and materials on the MELAB reading, and editing this working paper. My heartfelt appreciation is extended to the raters, the verbal report participants, and my colleague for their tremendous support.

References

Abbott, M. L. (2005). *English reading strategies: Differences in Arabic and Mandarin speaker performance on the CLBA reading assessment.* Unpublished doctoral dissertation, University of Alberta, Edmonton, Alberta, Canada.

Alderson, J. C. (1990). Testing reading comprehension skills (part one). *Reading in a Foreign Language*, *6,* 425–438.

Alderson, J. C. (2000). *Assessing reading.* Cambridge, UK: Cambridge University Press.

Alderson, J. C. (2005a, July). *The challenge of diagnostic testing: Do we know what we are measuring?* Plenary presented at the annual meeting of the Language Testing Research Colloquium, Ottawa, Canada.

Alderson, J. C. (2005b). *Diagnosing foreign language proficiency.* London: Continuum.

Alderson, J. C., & Lukmani, Y. (1989). Cognition and reading: Cognitive levels as embodied in test questions. *Reading in a Foreign Language*, *5,* 253–270.

Allan, A. (1992). *EFL reading comprehension test validation: Investigating aspects of process approaches.* Unpublished doctoral dissertation, Lancaster University, Lancaster, UK.

Anderson, R. C. (1972). How to construct performance tests to assess comprehension. *Review of Educational Research*, *42,* 145–170.

Anderson, N., Bachman, L., Perkins, K., & Cohen, A. (1991). An exploratory study into the construct validity of a reading comprehension test: Triangulation of data resources. *Language Testing*, *8*, 41–66.

Bachman, L. F. (1990). *Fundamental considerations in language testing.* Oxford: Oxford University Press.

Bachman, L. F. (2002). Some reflections on task-based language performance assessment. *Language Testing*, *19,* 453–476.

Bachman, L. F., Davidson, F., & Milanovic, M. (1996). The use of test method characteristics in the content analysis and design of EFL proficiency tests. *Language Testing, 13,* 125–150.

Bachman, L. F., Davidson, F., Ryan, K., & Choi, I. (1995). *An investigation into the comparability of two tests of English as a foreign language.* Cambridge, UK: Cambridge University Press.

Bachman, L. F., & Palmer, A. S. (1996). *Language testing in practice.* Oxford: Oxford University Press.

Bernhardt, E. (2003). Challenges to reading research from a multilingual world. *Reading Research Quarterly*, *38*, 112–117.

Block, E. (1992). See how they read: Comprehension monitoring of L1 and L2 readers. *TESOL Quarterly*, *26*, 319–341.

Breiman, L., Friedman, J. H., Olshen, R. A., & Stone, C. J. (1984). *Classification and regression trees*. Belmont, CA: Wadsworth International.

Butcher, K. R., & Kintsch, W. (2003). Text comprehension and discourse processing. In A. F. Healy, & R. W. Proctor (Eds.), *Handbook of psychology: Experimental psychology* (pp. 575–595). New York: John Wiley & Sons.

Carr, N. T. (2003). *An investigation into the structure of text characteristics and reader abilities in a test of second language reading*. Unpublished doctoral dissertation, University of California, Los Angeles, USA.

Carrell, P. L. (1983a). Some issues in studying the role of schemata, or background knowledge, in second language comprehension. *Reading in a Foreign Language*, *1*, 81–92.

Carrell, P. L. (1983b). Three components of background knowledge in reading comprehension. *Language Learning*, *33*, 183–203.

Carrell, P. L. (1984). The effects of rhetorical organization on ESL readers. *TESOL Quarterly*, *17*, 441–469.

Carrell, P. L. (1985). Facilitating ESL reading by teaching text structure. *TESOL Quarterly*, *19*, 727–752.

Carver, R. P. (1992). Effect of prediction activities, prior knowledge, and text type upon amount comprehended: Using rauding theory to critique schema theory research. *Reading Research Quarterly*, *27*, 165–174.

Cheng, L. (2003). *Academic reading strategies used by Chinese EFL learners: Five case studies*. Unpublished doctoral dissertation, University of British Columbia, Vancouver, British Columbia, Canada.

Cobb, T. (2004). Web VP (Version 2.0) [Computer software]. Montreal, Quebec, Canada: University of Montreal.

Cohen, A. D., & Upton, T. A. (2005, July). *Strategies in responding to the new TOEFL reading tasks*. Paper presented at the annual meeting of the Language Testing Research Colloquium, Ottawa, Ontario, Canada.

Crick, J. E., & Brennan, R. L. (1983). *Manual for GENOVA: A generalized analysis of variance system* [Computer software]. Iowa, IA: ACT.

DiBello, L. V., & Crone, C. (2001, April). *Technical methods underlying the PSAT/NMSQTTM enhanced score report*. Paper presented at the annual meeting of the National Council of Measurement in Education, Seattle, WA.

Douglas, D. (2000). *Assessing languages for specific purposes*. Cambridge: Cambridge University Press.

Douglas, D., & Hegelheimer, V. (2005, July). *Cognitive processes and use of knowledge in performing new TOEFL listening tasks*. Paper presented at the annual meeting of the Language Testing Research Colloquium (LTRC), Ottawa, Ontario, Canada.

Embretson, S. E. (1998). A cognitive design system approach to generating valid tests: Application to abstract reasoning. *Psychological Methods*, *3*, 380–396.

Embretson, S. E. (1999). Generating item during testing: Psychometric issues and models. *Psychometrika*, *64*, 407–433.

Embretson, S. E., & Gorin, J. S. (2001). Improving construct validity with cognitive psychology principles. *Journal of Educational Measurement, 38*, 343–368.

English Language Institute, University of Michigan. (2003). *Michigan English language assessment battery technical manual.*Ann Arbor, MI: English Language Institute, University of Michigan

Enright, M. K., Grabe, W., Koda, D., Mosenthal, P., Mulcahy-Ernt, P., & Schedl, M. (2000). *TOEFL 2000 reading framework: A working paper* (TOEFL Monograph Series MS-17). Princeton, NJ: Educational Testing Service.

Enright, M. K., Morley, M., & Sheehan, K. M. (2002). Items by design: The impact of systematic feature variation on item statistical characteristics. *Applied Measurement in Education*, *15*, 49–74.

Ericsson, K. A., & Simon, H. A. (1993). *Protocol analysis: Verbal reports as data.* Cambridge, MA: MIT Press.

Ewing, M., & Huff, K. (2004, April). *Using item difficulty modeling to evaluate skill categories.* Paper presented at the annual meeting of the National Council on Measurement in Education, San Diego, CA.

Farhady, H., & Hessamy, G. (2005, July). *An empirical investigation of the L2 reading comprehension skills.* Paper presented at the annual meeting of the Language Testing Research Colloquium, Ottawa, Ontario, Canada.

Freedle, R., & Kostin, I. (1993). The prediction of TOEFL reading comprehension item difficulty: Implications for construct validity. *Language Testing*, *10*, 131–170.

Freedle, R., & Kostin, I. (1999). Does the text matter in a multiple-choice test of comprehension? The case for the construct validity of TOEFL's minitalks. *Language Testing*, *16*, 2–32.

Gitomer, D. H., & Rock, D. (1993). Addressing process variables in test analysis. In N. Frederiksen, R. J. Mislevy, & I. I. Bejar (Eds.), *Test theory for a new generation of tests* (pp. 243–268). Hillsdale, NJ: Lawrence Erlbaum Associates.

Goodman, K. S. (1967). Reading: A psycholinguistic guessing game. *Journal of the Reading Specialist*, *6*, 126–135.

Gorin, J. S. (2002). *Cognitive and psychometric modeling of text-based reading comprehension GRE-V items.* Unpublished doctoral dissertation, University of Kansas, Lawrence, KS.

Grabe, W. (1991). Current development in second-language reading research. *TESOL Quarterly*, *25*, 375–406.

Grabe, W. (2002). Reading in a second language. In R. Kaplan (Ed.), *The Oxford handbook of applied linguistics* (pp. 49–59). New York: Oxford University Press.

Grabe, W., & Stoller, F. L. (2002). *Teaching and researching reading.* London: Pearson Education.

Hudson, T. (1996). *Assessing second language academic reading from a communicative competence perspective: Relevance for TOEFL 2000* (TOEFL Monograph Series MS-4). Princeton, NJ: Educational Testing Service.

Hudson, T. (1998). Theoretical perspectives on reading. *Annual Review of Applied Linguistics, 18*, 43–60.

Huff, K. (2003). *An item modeling approach to providing descriptive score reports.* Unpublished doctoral dissertation, University of Massachusetts, Amherst, MA.

Jamieson, J., Jones, S., Kirsch, I., Mosenthal, P., & Taylor, C. (2000). *TORFL 2000 framework: A working paper* (TOEFL Monograph Series MS-16). Princeton, NJ: Educational Testing Service.

Johnston, P. (1984). Prior knowledge and reading comprehension test bias. *Reading Research Quarterly, 21*, 220–239.

Kasai, M. (1997). *Application of the rule space model to the reading comprehension section of the Test of English as a Foreign Language (TOEFL).* Unpublished doctoral dissertation, University of Illinois at Urbana-Champaign, Urbana, IL.

Keppel, G., & Zedeck, S. (2001). *Data analysis for research designs: Analysis of variance and multiple regression/correlation approaches.* New York: W. H. Freeman & Company.

Kerlinger, F.N. (1979). *Behavioral research: A conceptual approach.* New York: Holt.

Kirsch, I. S., & Mosenthal, P. B. (1990). Exploring document literacy: Variables underlying the performance of young adults. *Reading Research Quarterly, 25,* 5-30.

Koda, K. (1996). L2 word recognition research: A critical review. *The Modern Language Journal*, *80*, 450–460.

Koda, K. (2005). *Insights into second language reading: A cross-linguistic approach.* New York: Cambridge University Press.

Kolen, M. J., & Brennan, R. L. (2004). *Test equating, scaling, and linking: Methods and practices.* New York: Springer.

Kolln, M. (1999). *Rhetorical grammar: Grammatical choices, rhetorical effects*. Needham Heights, MA: Allyn & Bacon.

Kostin, I. (2004). *Exploring item characteristics that are related to the difficulty of TOEFL dialogue items* (TOEFL Research Rep. No. RR-79). Princeton, NJ: Educational Testing Service.

LaBerge, D., & Samuels, S. J. (1974). Toward a theory of automatic information processing in reading. *Cognitive Psychology*, *6*, 293–323.

Lee, Y. (2004). Examining passage-related local item dependence (LID) and measurement construct using Q3 statistics in an EFL reading comprehension test. *Language Testing*, *21*, 74–100.

Leighton, J. P. (2004). Avoiding misconception, misuse, and missed opportunities: The collection of verbal reports in educational achievement testing. *Educational Measurement: Issues and Practice*, *23*(4), 6–15.

Leighton, J. P., & Gierl, M. (2005, May). *Identifying models of cognition in educational measurement.* Paper presented at the annual meeting of the Canadian Society for the Study of Education, London, Ontario, Canada.

Lumley, T., & Brown, A. (2004). Test-taker and rater perspectives on integrated reading and writing tasks in the Next Generation TOEFL. *Language Testing Update*, *35*, 75–79.

McKeown, M. G., Beck, I. L., Sinatra, G. M., & Losterman, J. A. (1992). The contribution of prior knowledge and coherent text to comprehension. *Reading Research Quarterly, 27*, 79–93.

Mislevy, R. J. (1994). Evidence and inference in educational assessment. *Psychometrika*, *59*, 439–483.

Mislevy, R. J. (1996). Test theory reconceived. *Journal of Educational Measurement, 33*, 379–416.

Mislevy, R. J., Almond, R. G., Yan, D., & Steinberg, L. S. (1999). Bayes nets in educational assessment: Where do the numbers come from? In K. B. Laskey, & H. Prade (Eds.), *Proceedings of the fifteenth conference on uncertainty in artificial intelligence* (pp. 437–446). San Francisco, CA: Morgan Kaufmann.

Mislevy, R. J., Steinberg, L. S., & Almond, R. G. (2002). Design and analysis in task-based language assessment. *Language Testing*, *19*, 477–496.

Munby, J. (1978). *A communicative syllabus design: A sociolinguistic model for defining the content of purpose-specific language programmes.* New York: Cambridge University Press.

National Research Council. (2001). *Knowing what students know: The science and design of educational assessment.* Washington, DC: National Academy Press.

Nelson, L. R. (2000). *Item analysis for tests and surveys using Lertap 5* [Computer software]. Perth, Australia: Curtin University of Technology.

Nichols, P. D. (1994). A framework for developing cognitively diagnostic assessments. *Review of Educational Research*, *64,* 575–603.

Perfetti, C. A. (1995). Cognitive research can inform reading education. *Journal of Research in Reading*, *18,* 106–115.

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.

Phillips, L. M., & Norris, S. P. (2002). Schema theory criticisms. In B. J. Guzzetti (Ed.), *Literacy in America: An encyclopedia of history, theory, and practice* (pp. 558–561). Santa Barbara, CA: ABC-CLIO.

Radford, A. (2004). *English Syntax: An introduction.* New York: Cambridge University Press.

Rayner, K., & Pollatsek, A. (1989). *The psychology of reading.* Englewood Cliffs, NJ: Prentice Hall.

Roller, C. (1990). Commentary: The interaction of knowledge and structure variables in the processing expository prose. *Reading Research Quarterly*, *25*, 79–89.

Ruddell, R. B., & Ruddell, M. R., & Singer, H. (1994). *Theoretical models and processes of reading.* Newark, Delaware: International Reading Association.

Rumelhart, D. E. (1977). Toward an interactive model of reading. In S. Domic (Ed.), *Attention and performance* (pp. 28–59). New York: Academic Press.

Rumelhart, D. E. (1980). Schemata: The building blocks of cognition. In R. J. Spiro, B. C. Bruce, & W. F. Brewer (Eds.), *Theoretical issues in reading comprehension: Perspectives from cognitive psychology, linguistics, artificial intelligence, and education* (pp. 33–58). Hillsdale, NJ: Lawrence Erlbaum Associates.

Rupp, A. A., Garcia, P., & Jamieson, J. (2001). Combining multiple regression and CART to understand difficulty in second language reading and listening comprehension tests. *International Journal of Testing*, *1*, 185–216.

Samejima, F. (1997). Graded response model. In W. J. Van der Linden & R. K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 85–100). Ann Arbor, MI: Edwards Brothers.

Sheehan, K. (1997). A tree-based approach to proficiency scaling and diagnostic assessment. *Journal of Educational Measurement, 34*, 333–352.

Sheehan, K., & Ginther, A. (2001, April). *What do multiple choice verbal reasoning items really measure? An analysis of the cognitive skills underlying performance on a standardized test of reading comprehension skill.* Paper presented at the annual meeting of the National Council on Measurement in Education, Seattle, WA.

Smith, F. (1971). *Understanding reading: A psycholinguistic analysis of reading and learning to read.* New York: Holt, Rinehart and Winston.

Smith, F. (2004). *Understanding reading: A psycholinguistic analysis of reading and learning to read*. Mahwah, NJ: Lawrence Erlbaum Associates.

SPSS for windows. (2005). *SPSS (Version 13.0)* [Computer software]. Chicago, IL: SPSS.

Stanovich, K. E. (1980). Towards an interactive compensatory model of individual differences in the development of reading fluency. *Reading Research Quarterly*, *16*, 32–71.

Stanovich, K. (2000). *Progress in understanding reading: Scientific foundations and new frontiers.* New York: Guilford Press.

Strong-Krause, D. (2001). *English as a second language speaking ability: A study in domain theory development.* Unpublished doctoral dissertation , Brigham Young University, Provo, UT.

Tatsuoka, K. K. (1995). Architecture of knowledge structures and cognitive diagnosis: A statistical pattern recognition and classification approach. In P. D. Nichols, S. F. Chipman, & R. L. Brennan (Eds.), *Cognitively diagnostic assessment* (pp. 327–359). Hillsdale, NJ: Erlbaum.

Thompson, G. (2004). *Introducing functional grammar* (2nd ed.). New York: Oxford University Press.

Thorndike, R. L. (1917). Reading as reasoning: A study of mistakes in paragraph reading. *Journal of Educational Psychological Association*, *8*, 323–332.

Urquhart, A. H., & Weir, C. J. (1998). *Reading in a second language: Process, product and practice.* New York: Addison Wesley Longman.

Wainer, H., & Lukhele, R. (1997). How reliable are TOEFL scores? *Educational & Psychological Measurement*, *57*, 741–758.

Wainer, H., Sheehan, K. M., & Wang, X. (2000). Some paths toward making Praxis scores more useful. *Journal of Educational Measurement*, *37*, 113–140.

Wang, X., Bradlow, E. T., & Wainer , H. (2002). A general Bayesian model for testlets: Theory and applications. *Applied Psychological Measurement*, *26*, 109–128.

Wang, X., Bradlow, E. T., & Wainer , H. (2004). *User's guide for SCORIGHT (version 3.0): A computer program for scoring tests built of testlets including a module for covariate analysis.* Princeton, NJ: Educational Testing Service; Philadelphia, PA: National Board of Medical Examiners.

Yang, P. (2000). *Effects of test-wiseness upon performance on the test of English as a foreign language.* Unpublished doctoral dissertation, University of Alberta, Edmonton, Alberta, Canada.

Appendix A

Consensus Codes for the MELAB Reading Items

Form/	Word	% Dif.			Pragmatic				
Item	Recog.	Words	Syntax	Text Org.	Knowledge	Locate	Inference	Synthesis	Distractor
E/1	2.00	19.10	2.00	2.00	.00	1.00	2.00	2.00	1.33
E/2	2.00	16.22	2.00	.00	.00	2.00	1.00	.00	3.00
E/3	2.00	21.22	1.00	2.00	4.00	.00	2.00	2.00	1.67
E/4	1.00	23.81	2.00	.00	3.00	1.00	1.00	.00	.67
E/5	2.00	28.36	2.00	2.00	3.00	1.00	2.00	1.00	2.00
E/6	2.00	18.18	.00	.00	.00	2.00	1.00	.00	2.33
E/7	1.00	24.00	1.00	.00	.00	1.00	1.00	.00	1.33
E/8	2.00	17.21	1.00	2.00	1.00	.00	2.00	2.00	2.33
E/9	.00	16.32	1.00	.00	.00	1.00	.00	.00	.00
E/10	.00	19.12	.00	.00	4.00	.00	3.00	.00	2.33
E/11	2.00	23.47	1.00	.00	3.00	1.00	.00	.00	.67
E/12	1.00	25.00	2.00	.00	1.00	2.00	1.00	.00	2.00
E/13	1.00	41.67	1.00	2.00	3.00	1.00	.00	1.00	.00
E/14	2.00	31.92	2.00	2.00	2.00	1.00	1.00	1.00	1.33
E/15	1.00	32.00	2.00	.00	4.00	2.00	1.00	.00	2.00
E/16	2.00	25.81	2.00	1.00	3.00	2.00	1.00	.00	2.00
E/17	1.00	22.44	1.00	.00	3.00	2.00	1.00	1.00	2.00
E/18	.00	10.00	2.00	1.00	3.00	2.00	1.00	1.00	1.00
E/19	1.00	12.87	2.00	2.00	.00	1.00	2.00	2.00	2.00
E/20	1.00	12.12	2.00	.00	3.00	1.00	1.00	1.00	1.67
F/1	1.00	27.78	1.00	1.00	.00	2.00	.00	1.00	.00
F/2	2.00	20.37	2.00	1.00	.00	1.00	1.00	1.00	1.00
F/3	1.00	12.50	2.00	.00	.00	2.00	.00	.00	1.33
F/4	1.00	31.58	1.00	.00	3.00	2.00	1.00	.00	2.67
F/5	1.00	25.00	1.00	.00	2.00	1.00	.00	1.00	1.00
F/6	1.00	22.61	1.00	2.00	.00	.00	2.00	2.00	2.33
F/7	1.00	20.31	1.00	1.00	3.00	1.00	2.00	1.00	1.67
F/8	2.00	25.58	2.00	2.00	4.00	2.00	1.00	1.00	1.00
F/9	2.00	19.04	2.00	2.00	4.00	1.00	1.00	2.00	1.00
F/10	2.00	20.31	2.00	1.00	3.00	1.00	1.00	1.00	1.00
F/11	1.00	19.48	1.00	2.00	2.00	.00	2.00	2.00	1.33
F/12	1.00	22.22	2.00	1.00	3.00	1.00	2.00	1.00	3.00
F/13	2.00	17.64	2.00	1.00	3.00	2.00	2.00	2.00	3.00
F/14	2.00	23.81	2.00	2.00	3.00	1.00	2.00	2.00	2.33
F/15	2.00	18.19	2.00	1.00	3.00	1.00	2.00	2.00	2.33
F/16	1.00	13.23	1.00	1.00	2.00	.00	1.00	2.00	.33
F/17	1.00	10.00	2.00	.00	3.00	2.00	.00	.00	1.67
F/18	2.00	25.00	2.00	1.00	1.00	2.00	2.00	1.00	1.67
F/19	2.00	13.80	1.00	2.00	1.00	1.00	2.00	2.00	1.33
F/20	2.00	8.70	2.00	.00	.00	1.00	1.00	1.00	1.67

Appendix B

Item Difficulty Parameter Estimates for the Reading Items on Form E and Form F

Form / Item	Item Difficulty (*b*)	Form / Item	Item Difficulty (*b*)
E1	.26	F1	-.71
E2	-.30	F2	-.62
E3	-.62	F3	-.42
E4	-.72	F4	1.11
E5	1.24	F5	-1.16
E6	-.25	F6	.47
E7	-.46	F7	.84
E8	.85	F8	-.86
E9	-1.16	F9	.09
E10	1.54	F10	-.45
E11	-.23	F11	.43
E12	.09	F12	1.39
E13	-1.01	F13	1.48
E14	.47	F14	1.11
E15	1.31	F15	1.78
E16	.65	F16	-1.03
E17	1.88	F17	1.37
E18	.19	F18	1.57
E19	.67	F19	.54
E20	.23	F20	1.53

Appendix C

Comparison of Item Coding and Actual Processes Involved in Correctly Answering Each Item

Form/	Word	Syntax	Text	Pragmatic	Locate	Inference	Synthesis	Distractor
Item	Recog.		Org.	Knowledge				
E/1	*2.00	*2.00	2.00	0.00	*1.00	*2.00	*2.00	*1.33
E/2	*2.00	2.00	*0.00	0.00	*2.00	1.00	*0.00	*3.00
E/3	*2.00	1.00	*2.00	4.00	*0.00	*2.00	*2.00	*1.67
E/4	1.00	2.00	*0.00	*3.00	*1.00	*1.00	*0.00	*0.67
E/5	2.00	2.00	*2.00	3.00	*1.00	*2.00	1.00	2.00
E/6	*2.00	*0.00	*0.00	0.00	*2.00	1.00	*0.00	*2.33
E/7	1.00	*1.00	*0.00	0.00	*1.00	1.00	*0.00	1.33
E/8	2.00	1.00	*2.00	*1.00	*0.00	*2.00	*2.00	*2.33
E/9	*0.00	1.00	*0.00	*0.00	*1.00	*0.00	*0.00	*0.00
E/10	*0.00	*0.00	*0.00	*4.00	*0.00	*3.00	*0.00	*2.33
E/11	*2.00	*1.00	*0.00	3.00	*1.00	*0.00	*0.00	*0.67
E/12	1.00	*2.00	*0.00	1.00	*2.00	1.00	*0.00	*2.00
E/13	*1.00	1.00	*2.00	3.00	*1.00	*0.00	*1.00	*0.00
E/14	*2.00	2.00	*2.00	2.00	*1.00	*1.00	1.00	*1.33
E/15	*1.00	*2.00	*0.00	4.00	*2.00	*1.00	*0.00	2.00
E/16	*2.00	2.00	1.00	3.00	*2.00	*1.00	*0.00	*2.00
E/17	*1.00	1.00	*0.00	3.00	*2.00	*1.00	*1.00	*2.00
E/18	*0.00	*2.00	*1.00	3.00	*2.00	*1.00	*1.00	*1.00
E/19	*1.00	*2.00	2.00	*0.00	*1.00	*2.00	*2.00	*2.00
E/20	1.00	2.00	*0.00	3.00	*1.00	*1.00	*1.00	*1.67
F/1	*1.00	*1.00	1.00	0.00	*2.00	*0.00	*1.00	*0.00
F/2	*2.00	2.00	1.00	0.00	*1.00	*1.00	*1.00	*1.00
F/3	*1.00	*2.00	*0.00	0.00	*2.00	*0.00	*0.00	1.33
F/4	*1.00	*1.00	*0.00	3.00	*2.00	*1.00	*0.00	2.67
F/5	*1.00	*1.00	*0.00	2.00	*1.00	*0.00	*1.00	*1.00
F/6	*1.00	1.00	2.00	*0.00	*0.00	*2.00	*2.00	*2.33
F/7	1.00	1.00	1.00	3.00	*1.00	*2.00	1.00	*1.67
F/8	*2.00	2.00	2.00	4.00	*2.00	*1.00	*1.00	*1.00
F/9	2.00	*2.00	*2.00	4.00	*1.00	*1.00	*2.00	*1.00
F/10	*2.00	*2.00	1.00	3.00	*1.00	*1.00	1.00	*1.00
F/11	1.00	1.00	*2.00	*2.00	*0.00	*2.00	*2.00	*1.33
F/12	*1.00	*2.00	*1.00	3.00	*1.00	*2.00	1.00	*3.00
F/13	*2.00	2.00	1.00	3.00	*2.00	*2.00	2.00	*3.00
F/14	2.00	*2.00	2.00	3.00	*1.00	*2.00	*2.00	*2.33
F/15	*2.00	*2.00	1.00	*3.00	*1.00	2.00	*2.00	*2.33
F/16	1.00	1.00	*1.00	*2.00	*0.00	1.00	*2.00	*0.33
F/17	1.00	*2.00	*0.00	3.00	*2.00	*0.00	*0.00	*1.67
F/18	*2.00	*2.00	*1.00	1.00	*2.00	*2.00	*1.00	*1.67
F/19	2.00	1.00	2.00	*1.00	*1.00	*2.00	*2.00	*1.33
F/20	*2.00	*2.00	*0.00	0.00	*1.00	*1.00	1.00	*1.67

* Cognitive processes reported by the participants who correctly answered the item.