CaMLA

CaMLA Working Papers
2015–02

The Characteristics of the Michigan English Test Reading Texts and Items and their Relationship to Item Difficulty

Khaled Barkaoui
York University
Canada

The Characteristics of the Michigan English Test Reading Texts and Items and their Relationship to Item Difficulty

Author

Khaled Barkaoui
Faculty of Education
York University, Canada

About the Author

Khaled Barkaoui is an associate professor at the Faculty of Education, York University, Canada. His current research and teaching focus on L2 assessment, L2 writing, L2 program evaluation, longitudinal research, mixed-methods research, and EAP. His publications have appeared in *Applied Linguistics*, *Assessing Writing*, *Language Testing*, *Language Assessment Quarterly*, *System*, and *TESOL Quarterly*. In 2012, he received the TOEFL Outstanding Young Scholar Award in recognition of the outstanding contributions his scholarship and professional activities have made to the field of second language assessment.

Table of Contents

Abstract

This study aimed, first, to describe the linguistic and discourse characteristics of the Michigan English Test (MET) reading texts and items and, second, to examine the relationships between the characteristics of MET texts and items, on the one hand, and item difficulty and bias indices, on the other. The study included 54 reading texts and 216 items from six MET forms that were administered to 6,250 test takers. The MET texts and items were coded in terms of 22 features. Next, item difficulty and bias indices were estimated. Then, the relationships between the characteristics of MET reading texts and items, on the one hand, and item difficulty and bias indices, on the other, were examined. The findings indicated that the sample of MET texts and items included in the study exhibited several desirable features that support the validity argument of the MET reading subsection. Additionally, some problematic characteristics of the texts and items were identified that need to be addressed in order to improve the test. The study demonstrates how to combine task and score analyses in order to examine important questions concerning the validity argument of second-language reading tests and to provide information for improving texts and items on such tests.

This study aimed, first, to describe the linguistic and discourse characteristics of the reading texts and items of the Michigan English Test (MET) and, second, to examine the relationships between the characteristics of these texts and items, on the one hand, and item difficulty, on the other. Typically, validation studies of reading comprehension tests examine the factor structure of such tests or the psychometric properties of their items (e.g., item difficulty, item discrimination). However, as Gorin and Embreston (2006) argued, examining the patterns of relationships among test scores (e.g., factor analysis) primarily supports "the *significance* of a construct, rather than its *meaning*" (p. 395, emphasis added). Gorin and Embreston maintained that it is the analysis of item and text characteristics (i.e., task analysis) and the examination of their relationships with item psychometric properties (e.g., item difficulty) that can "provide important information regarding *the substantive meaning of the construct* underlying questions in reading comprehension tests" (p. 395, emphasis added; cf. Alderson, 2000; Khalifa & Weir, 2009).

Studies that aim to describe the characteristics of reading texts and items and to examine their relationships with the difficulty indices of reading comprehension items typically involve three stages (e.g., Gorin & Embreston, 2006; In'nami & Koizumi, 2009; Ozuru, Rowe, O'Reilly, & McNamara, 2008; Rupp, Garcia, & Jamieson, 2001). First, the text and item characteristics deemed important based on test specifications and/or theory and research on reading comprehension and item response processes (e.g., item format, text length) are identified and coded. Second, the difficulties of test items are estimated through the statistical analysis of item scores. Third, the relationships between item and text characteristics and item difficulty are examined. The result is a detailed description of item and text characteristics and a list of item and text factors that contribute to variability in item psychometric properties and, by extension, variability in performance on reading comprehension tests (e.g., In'nami & Koizumi, 2009; Ozuru et al., 2008; Rupp et al., 2001). This line of research can provide important validity evidence and contribute useful information for the development and improvement of reading comprehension test specifications, tasks, and items (Alderson, 2000; Buck, Tatsuoka, & Kostin, 1997; Dávid, 2007; Embretson & Wetzel, 1987; Freedle & Kostin, 1993; Gorin & Embreston, 2006; In'nami & Koizumi, 2009; Khalifa & Weir, 2009; Ozuru et al., 2008; Rupp et al., 2001; Spelberg, de Boer, & van den Bos, 2000).

Task Analysis as Validity Evidence

Analysing the characteristics of items and texts on reading comprehension tests (i.e., task analysis) can provide important information about the meaning of the construct(s) of such tests (Alderson, 2000; Buck et al., 1997; Embretson & Wetzel, 1987; Freedle & Kostin, 1993; Gao, 2006; Gorin & Embreston, 2006; Khalifa & Weir, 2009; Kirsch & Guthrie, 1980; Ozuru et al., 2008; Rupp et al., 2001). Kirsch and Guthrie (1980), for example, maintained that descriptions of item and text variables on reading tests help to describe the constructs actually measured by such tests as well as the factors that are likely to influence variation in test performance (cf. Gorin & Embreston, 2006). Building on research on text processing and comprehension, many studies have decomposed the process of responding to items on reading comprehension tests into a processing model and then analyzed the characteristics of reading comprehension texts and items in terms of their necessary cognitive processes in order to identify the contribution of these processes to item responses (e.g., Davey, 1988; Embretson & Wetzel, 1987; Freedle & Kostin, 1993; Gorin & Embretson, 2006; Ozuru et al., 2008). The results of these analyses can provide significant insights into the factors (e.g., item and text characteristics) and the types of cognitive processing (e.g., vocabulary knowledge, syntactic processing) that are involved in responding to reading comprehension items.

One study serves to illustrate the use of findings from theory and research on reading comprehension and item response processes to identify relevant item and text features that relate to those processes. Embretson and Wetzel (1987) analyzed multiple-choice items on a first-language (L1) reading comprehension test using a coding system based on a cognitive processing model of reading comprehension (see also Davey, 1988). According to this model, answering multiple-choice reading comprehension questions involves two stages: text representation and response decision. The first stage involves encoding the reading text by forming a coherent representation of its overall meaning that integrates text-based information and prior knowledge. Coherence is the process of connecting word meanings and propositions into a meaningful representation of the text. The difficulty of encoding is controlled by linguistic features of the text, particularly vocabulary difficulty, while the difficulty of coherence processes is most strongly influenced by the propositional density of the text. Texts with more difficult vocabulary are more difficult to encode and consequently more difficult to retrieve when responding to comprehension questions. Propositionally dense texts are difficult to process and integrate for later recall and comprehension because of working memory capacity limitations that preclude holding large amounts of information simultaneously.

The second stage of Embretson and Wetzel's (1987) model, response decision, involves three steps. The first step is encoding, which involves forming a representation of the meaning of the stem and the response alternatives. The second step is mapping, which involves relating the propositions in the stem and response alternatives to the information retrieved from the text. Finally, evaluating the truth status of the response alternatives involves a two-stage process of falsification and confirmation of response alternatives. The two decision processes of text representation and response decision describe the extent to which information given in the text could be used to make decisions regarding the response options. For example, items with correct responses that are directly confirmable by information in the text or distractors that are explicitly contradicted by the text require little processing.

According to Embretson and Wetzel's (1987) model, item difficulty is influenced by the difficulty of the processing required in each of the components as well as by item and text factors. Thus, an item is expected to be more difficult when the text contains more information; the item requests more information; overlap between the answer options and the text content is small; and/or answer options cannot be explicitly confirmed by the text content. Difficulty in text mapping is partially influenced by the amount of information needed from the text to answer the question. As the amount of text relevant to answering a question increases, so do the demands on memory, encoding, and item difficulty. Embretson and Wetzel coded reading comprehension items and texts in their study in terms of features that were theoretically related to the processing components identified in their model and then examined the relationships between these features and item difficulty. The findings of this study, as well as other relevant studies, are discussed below.

Examining the Relationships between Item and Text Features and Item Difficulty

Examining the relationships between item and text factors, on the one hand, and item difficulty, on the other, can provide important validity evidence by identifying and estimating the contribution of construct-relevant and construct-irrelevant factors to variability in item difficulty and test performance (Buck et al., 1997; Freedle & Kostin, 1993; Gao, 2006; Gorin & Embreston, 2006; Kirsch & Guthrie, 1980; Rupp et al., 2001; Spelberg et al., 2000). As Messick (1989) explained, the largest proportion of variance in test scores (and, by extension, item difficulty) should be construct-relevant, i.e., reflect what the test intends to measure. Comparatively little score variance should be construct-irrelevant, i.e., contributed by factors other than the construct being measured. Construct-irrelevant variance may derive from different sources including test method. Alderson (2000), for example, noted that, depending on the purpose and uses of a test, the effects of some text and item factors might be desirable, but "others might be irrelevant to what is supposedly being tested" (p. 90). For instance, if the correct option in a multiple-choice item has low-frequency vocabulary and/or the distractors have different levels of falsifiability, and this affects item difficulty, then this constitutes a method effect that is construct-irrelevant. The falsifiability of distractors is construct-irrelevant because eliminating improbable answer options is associated more strongly with general test-taking skills than with reading comprehension skills (Ozuru et al., 2008). Alderson emphasized that it is crucial that what a test measures is "as little contaminated as possible by the test method" (p. 115; cf. Gorin & Embreston, 2006; Khalifa & Weir, 2009).

Freedle and Kostin (1993) explained that a key validity question for reading comprehension tests is whether answering items on such tests is influenced more by the characteristics of the item itself or by the content and structure of the text to be comprehended. Since the purpose of reading comprehension tests is to assess whether the text itself has been comprehended, if item factors (e.g., item format, item vocabulary level) have a greater impact on test performance than do text characteristics (e.g., text abstractness, rhetorical organization), then one cannot claim that the test is construct valid (*sic*). In other words, if item difficulty has a higher correlation with item variables (e.g., item vocabulary familiarity) than it does with text variables (e.g., text vocabulary difficulty), then this is evidence that the items fail to capture comprehension skills related directly to the texts associated with the questions (cf. Kirsch & Guthrie, 1980). This weakens the validity of the test as a measure of reading comprehension. By contrast, if the difficulty of reading comprehension items is determined primarily by those text and text-related variables that theory and research have shown to influence comprehension processes then this constitutes evidence that the test is in fact a measure of text comprehension (cf. Alderson, 2000; Gorin & Embreston, 2006; Khalifa & Weir, 2009; Ozuru et al., 2008; Spelberg et al., 2000).

Green (1984) cautioned that the analysis of item difficulty alone cannot answer the question of whether and to what extent a test is valid because items measuring the same construct may, and generally do, differ in difficulty. By analysing and comparing items that differ in difficulty, "inferences may be made about the demands on processing and knowledge" required for answering reading comprehension items correctly (p. 552). Similarly, Dávid (2007) argued that while item difficulty does not in itself provide useful information about validity, construct-irrelevant difficulty or easiness can weaken the validity of score-based inferences. In other words, while item difficulty that springs from the focus of the item (e.g., specific details or main idea) is considered to be relevant to the construct, difficulty that comes from the method (e.g., the format of the item) or other sources (e.g., unclear instructions) is construct-irrelevant. While score analyses can identify which items are more or less difficult, they cannot explain why. Examining the relationships between item and text characteristics and indices of item difficulty can provide such an explanation (cf. Gorin & Embreston, 2006; Ozuru et al., 2008; Rupp et al., 2001).[1] Such research also allows test developers to better define the constructs that they are testing (Gorin & Embreston, 2006). In particular, it can help reveal the key processing variables that relate to item difficulty and identify the features that account for comprehension item processing; information that is important for establishing the validity of score-based inferences about test-taker reading ability (Gorin & Embreston, 2006).

Previous Studies

Several studies have examined the relationships between text and item features on the one hand and item difficulty on the other in L1 reading tests (e.g., Davey, 1988; Davey & Lasasso, 1984; Embretson & Wetzel, 1987; Gorin & Embreston, 2006; Green, 1984; Hare, Rabinowitz, & Schieble, 1989; Kintsch & Yarbough, 1982; Kirsch & Guthrie, 1980; Ozuru et al., 2008; Rupp et al., 2001) and L2 reading tests (e.g., Alderson et al., 2006; Bachman, Davidson, Ryan, & Choi, 1995; Freedle & Kostin, 1993; Gao, 2006; Kobayashi, 2002). Some of the text factors that these studies have

[1] Think-aloud protocols of reading processes while completing a reading test can also provide useful information for explaining reading item difficulty (cf. Anderson, Bachman, Perkins, & Cohen, 1991; Cohen & Upton, 2007; Gao, 2006).

examined include: length, topic, linguistic characteristics (e.g., sentence complexity, vocabulary level), rhetorical organization, coherence, concreteness/abstractness, readability, and propositional density. Reading comprehension research indicates that these factors can influence reading comprehension processes significantly by affecting the text representation component of the processing model. For example, longer texts and texts with more complex syntactic structures, longer sentences, more referential expressions, specialized and less frequent vocabulary, higher propositional density, and higher level of abstractness are more difficult to process, understand, and recall when answering comprehension questions, resulting in lower performance on the questions associated with these texts (Embretson & Wetzel, 1987; Freedle & Kostin, 1993; Gorin & Embretson, 2006; Ozuru et al., 2008).

Some of the item variables that have been examined in previous studies include: item language (i.e., syntactic complexity and vocabulary level), item format (e.g., multiple-choice, short answer, matching), item focus (e.g., main idea, specific details), and inference type required. For instance, questions that require test takers to engage in inferential processes are likely to be harder than those that require simple matching of question and text, while questions with low-frequency vocabulary can present test takers with an additional layer of difficulty (Alderson, 2000). Most of the studies cited above focused on various features of multiple-choice questions such as stem length and vocabulary level, length and vocabulary level of response options, degree of lexical similarity/overlap between the correct answer and the distractors, and the number of explicitly falsifiable distractors. For example, Embretson and Wetzel (1987) and Gorin and Embreston (2006) found that the vocabulary level of the correct response and the distractors influenced item difficulty. Rupp et al. (2001) found that items with highly similar options in terms of wording (i.e., lexical overlap) were harder than items with options that have lower overlap. Additionally, items with longer sentences and higher type-token ratios (a measure of lexical diversity) were more difficult than shorter items with low type-token ratios (Rupp et al., 2001).

Several of the studies cited above examined item-by-text interaction effects on item difficulty as well. Item-by-text interaction effects involve an interaction between what the test taker is required to do and the content and structure of the text (Rupp et al., 2001). Examples of item-by-text variables examined in previous studies include: whether the requested information is implicitly or explicitly mentioned in the text, the location of requested information in the text, the proportion of text that is relevant to the question to be answered, the level of abstractness or concreteness of the information requested by the question, and lexical overlap between item options and text. Generally, questions requiring the synthesis of information from various locations in the text are harder than questions referring to information in one location only; questions where there is lower lexical overlap between the text and the question are harder than questions with greater overlap; items requiring implicit information in the text are more difficult than items requiring explicit information; and items requiring more abstract information or more inferences are more difficult (Alderson, 2000; Davey, 1988; Davey & Lasasso, 1984; Freedle & Fellbaum, 1987; Ozuru et al., 2008; Rupp et al., 2001).

Previous studies on the relationships between item and text features and item psychometric properties tended to examine only one index of item quality, item difficulty. Few studies have examined other indices such as item discrimination. Most of these studies estimated item difficulty and discrimination using classical test theory (CTT) procedures. CTT, however, has its limitations. In particular, CTT estimates of item difficulty lack stability as they are dependent on the sample of test takers who answer the items. In contrast, Item Response Theory models, including Rasch models, provide item statistics that are independent of the groups from which they are estimated. The result is stable estimations of item difficulty on a true interval scale (Barkaoui, 2013a; Bond & Fox, 2007; McNamara, 1996).

Additionally, Rasch analyses allow the examination of other aspects of item quality that cannot be examined using CTT analyses such as item fit and item bias. Fit statistics provide information about the extent to which the response data fit; that is, perform according to the Rasch Model. As such, fit statistics are an important indicator of item quality (Barkaoui, 2013a; Bond & Fox, 2007; McNamara, 1996). Bias analysis investigates whether a particular aspect of the assessment setting elicits a consistently biased pattern of scores. Bias analysis is similar to Differential Item Functioning (*DIF*) analysis in that it aims to identify any systematic subpatterns of behavior occurring from an interaction of particular items with particular subgroups of test takers and to estimate the effects of these interactions on test scores (Barkaoui, 2013a; Bond & Fox, 2007; McNamara, 1996; Uiterwijk & Vallen, 2005). Examining the relationships between item and text factors and item bias can provide important information about test validity (Davey, 1988; Spelberg et al., 2000; Uiterwijk & Vallen, 2005). As Uiterwijk and Vallen (2005) explained, merely detecting item bias does not identify the element in the item that causes it. One way to identify sources of item bias is to examine the characteristics of items and texts in the reading test that show bias and compare them to those of other items. The current study examined three indicators of item quality for the MET reading subsection using a Rasch model: item difficulty, item fit, and item bias.

The Present Study

This study aimed, first, to describe the linguistic and discourse characteristics of the MET reading texts and items and, second, to examine the relationships between the characteristics of MET texts and items on the one hand and item difficulty, fit, and bias indices on the other. According to CaMLA (2012), the MET reading subsection aims “to assess the comprehension of a variety of written texts in social, educational, and workplace contexts” (p. 7). This definition puts emphasis on text comprehension as the construct being assessed. From this perspective, item-related factors should *not* contribute to variability in test performance and item difficulty. Evidence that text and text-by-item factors are the main contributors to variability in item difficulty supports the test’s validity argument. Additionally, the study aimed to identify item- and text-related sources of item bias and misfit, if any. The study addressed the following research questions:

1. What are the linguistic and discourse characteristics of a sample of MET reading texts and items?
2. What are the difficulties and fit of a sample of MET reading items?
3. To what extent do item, text, and item-by-text variables relate to item difficulty and fit in the MET reading subsection?
4. Are there any biased interactions between test-taker subgroups (i.e., test-taker age, gender, L1) and items in the MET reading subsection?
5. What item, text, and item-by-text variables, if any, distinguish biased and nonbiased items in the MET reading subsection?

The Michigan English Test (MET) Reading Subsection

The Michigan English Test (MET) is an international, standardized, English as a foreign language (EFL), multi-level examination designed by CaMLA (Cambridge Michigan Language Assessments) and intended for adults and adolescents at or above a secondary level of education. The MET targets a range of proficiency levels from upper beginner to lower advanced levels (levels A2 to C1 of the Common European Framework of Reference, CEFR), with emphasis on the middle of the range (i.e., levels B1 and B2 of CEFR) (CaMLA, 2012). The MET emphasizes the ability to communicate effectively in English in a variety of linguistic contexts and assesses test takers’ English language proficiency in three language skill areas: listening, reading, and language usage (grammar and vocabulary). The MET is a paper-and-pencil test that includes 135 multiple-choice questions in two sections: (a) Listening and (b) Reading and Grammar. Listening recordings and reading passages focus on three domains or contexts (public, educational, and occupational) and reflect a range of topics and situations. The MET is administered every month at authorized test centers around the world with a new test form developed for each administration.

The MET reading subsection aims to assess the test-taker's ability to understand a variety of written texts in social, educational, and workplace contexts (CaMLA, 2012). Each form of the test consists of three reading sets; each set includes three reading texts and 12-14 multiple-choice questions. The three texts in each set are thematically linked but each text belongs to one of three different text types, called sections A, B and C. Section-A texts are typically about 80-word long and consist of a short message, announcement, advertisement, description, or other type of text typical of those found in newspapers and newsletters. Section-B texts are about 160-word long and consist of a short text such as a segment of a glossary, a memo, a letter to the editor, or a resume. Section-C texts are longer (about 290 words or 3 to 5 paragraphs) and more abstract than texts in sections A and B; they typically consist of an academic article that includes argument or exposition. Each question has four options and one correct answer. Typically, there are between two and five items per text. Additionally, one or two questions in each set require test takers to synthesize information presented in two or three texts (CaMLA, 2012). The reading texts and items on the MET reflect a range of

Table 1: Numbers of Forms, Texts, Items and Test Takers Included in the Study

MET Form	Number of Test Takers	Number of Texts	Number of Items
1	1,195	9	36
2	932	9	36
3	779	9	36
4	1,325	9	36
5	996	9	36
6	1,023	9	36
Total	6,250	54	216

situations that occur in three domains: public spaces (e.g., street, shops, restaurants, sports or entertainment) and other social networks outside the home; occupational workplace settings (e.g., offices, workshops, conferences); or educational settings (e.g., schools, colleges, classrooms, residence halls). The texts cover a wide range of topics that do not require any specialized knowledge or experience to understand. Each reading set assesses three reading subskills: global comprehension (e.g., understanding main idea; identifying speaker's purpose; synthesizing ideas from different parts of the text), local comprehension (e.g., identifying supporting detail; understanding vocabulary; synthesizing details; recognizing restatement), and inferencing (e.g., understanding rhetorical function; making an inference; inferring supporting detail; understanding pragmatic implications) (CaMLA, 2012). MET reading items are scored automatically by computer.

Methods

Dataset

The study included a sample of six forms of the MET reading subsection and a sample of 6,250 MET test takers who responded to these six forms (see Table 1). As Table 1 shows, each MET form included nine reading texts and 36 multiple-choice questions, for a total of 54 reading texts and 216 items. The number of test takers who responded to each form varied between 779 (form 3) and 1,325 (form 4).

Table 2 displays descriptive statistics concerning the demographics of the test takers included in the study by test form. Slightly more than half the test takers (56.3%) were females. The test-taker's ages ranged between 11 and 62 years, with the majority (87%) being between 11 and 30 years. The test takers spoke nine different first languages (L1), with the great majority being L1 speakers of Spanish (84.5%), followed by Albanian (14.4%) and Other (1.1%). The distribution of the

Table 2: Test Takers' Demographics

MET Form	Percentage Females	Age Range Percentage					L1 Percentage		
		11–18	19–24	25–30	31–40	Over 40	Spanish	Albanian	Other
1	60.4	9.9	51.1	25.8	9.1	4.1	72.4	23.0	4.6
2	60.4	10.4	48.9	26.4	10.0	4.3	78.3	21.5	0.2
3	57.4	22.8	44.1	19.5	10.5	3.1	99.5	0.0	0.5
4	55.1	40.7	32.3	15.7	6.7	4.6	86.5	12.6	0.9
5	53.2	22.2	46.0	18.8	7.9	5.2	82.8	17.0	0.2
6	51.1	22.8	48.4	18.1	7.5	3.3	87.7	12.3	0.0
Total	56.3	21.5	45.1	20.7	8.6	4.1	84.5	14.4	1.1

test takers included in this study in terms of age and gender seems similar to that of MET test takers in 2013 (CaMLA, 2014).

Data Analysis

Data analysis consisted of three phases. First, a detailed analysis was conducted of the linguistic and discourse characteristics of the sample of MET reading texts and items included in the study to address research question 1. Second, item difficulty, fit, and bias indices were estimated to address research questions 2 and 4. Third, the relationships between the characteristics of MET reading texts and items on the one hand and item difficulty and bias indices on the other were examined to address research questions 3 and 5.

Coding item and text characteristics

Each reading text and item in each MET form was coded in terms of various text, item, and item-by-text features as listed in Table 3. These features were selected based on reviews of (a) the features used by CaMLA (CaMLA, 2012) to describe MET reading texts and items (e.g., text domain, subskill tested by item) and (b) features shown by theory and research to affect reading comprehension processes and reading item difficulty (e.g., Alderson, 2000; Alderson et al., 2006; Bachman et al., 1995; Embretson & Wetzel, 1987; Enright et al., 2000; Freedle & Kostin, 1993; Gorin & Embreston, 2006; Khalifa & Weir, 2009; Rupp et al., 2001).

MET reading texts and items were coded using computer programs and manually. The main computer program used to analyse the texts and items in this study was *Coh-Metrix*, a web-based software that provides more than 100 indices of cohesion, vocabulary, syntactic complexity, and text readability; features that have been show to influence reading comprehension (Crossley, Greenfield, & McNamara, 2008; Crossley, Louwerse, McCarthy, & McNamara, 2007; Crossley & McNamara, 2008; Graesser, McNamara, Louwerse, & Cai, 2004; Green, Unaldi, & Weir, 2010; Ozuru et al., 2008). *Coh-Metrix* represents an advance on conventional readability measures because it allows the examination of various linguistic and discourse features (e.g., lexical and syntactic features, cohesion and coherence) that are related to text processing and reading comprehension (Crossley et al., 2007, 2008; Crossley & McNamara, 2008; Graesser et al., 2004). Crossley et al. (2007), for example, used *Coh-Metrix* to compare the linguistic and discourse characteristics of simplified and authentic texts used in ESL reading textbooks, while Green et al. (2010) used it to analyze and compare test and authentic reading texts in terms of various features that affect text difficulty in order to assess the authenticity of test reading texts.

For manually coded variables in Table 3, two researchers, both graduate students in applied linguistics, were trained on the coding scheme and then independently coded all the MET reading texts and items in terms of various text and item-by-text features. The codings were then compared and inter-coder agreement percentage (i.e., number of agreements divided by total number of decisions) was computed for each coded feature (see Table 3 for inter-coder agreement percentage for each manually-coded variable). Disagreements for each manually-coded feature were discussed, resolved, and then a final code was assigned to each feature. The following paragraphs describe each of the variables listed in Table 3.

Text Variables

As Table 3 shows, the study included 10 sets of variables related to text characteristics. Variables 1 and 2 (section and domain) were based on information from CaMLA (2012). Each text is classified by CaMLA as belonging to one of three text types, called sections: A, B or C. For domain, each of the 54 texts in the study was classified as belonging to one of three domains: public, occupational, or educational. Variable 3, topic, concerns the subject matter of the text. Each text was classified as being on one of five topics: health and psychology, environment, economics and job-related, science and technology (including computer, communication, and transportation), or everyday life (e.g., entertainment, food, leisure, tourism, arts). Because some texts included nonverbal information (e.g., tables, illustrations, pictures, drawings), each text was also coded as including nonverbal information (coded 1) or not (coded 0) (i.e., variable 4 in Table 3).[2]

The remaining six text features in Table 3 (variables 5 to 10) were all estimated using *Coh-Metrix*. All these variables have been shown by previous research to affect reading comprehension. *Length* refers to the number of words in the text. Generally, longer texts require more processing and have higher memory load and integration requirements than shorter texts (Gorin & Embreston, 2006; Rupp et al., 2001). *Syntactic Complexity* refers

[2] It should be noted here that, according to the test developers, the illustrations in the MET texts are intended to be decorative, rather than informational.

Table 3: Item, Text, and Item-by-Text Variables Examined in the Study

	Variables
Text	1. Section
	2. Domain (83%)*
	3. Topic (84%)
	4. Nonverbal information (100%)
	5. Length
	6. Syntactic complexity: Sentence length and syntactic similarity
	7. Lexical Features: Lexical density, lexical variation (MTLD), lexical sophistication (lambda, AWL), and word information (word frequency, familiarity, and polysemy).
	8. Coherence and Cohesion: Referential cohesion (Argument overlap between adjacent sentences), conceptual cohesion (LSA mean all sentences similarity), and connectives density.
	9. Text Concreteness/Abstractness: *Coh-Metrix* z-score for word concreteness
	10. Text readability: Flesch Reading Ease
Item	11. Item Length
	12. Item vocabulary: word familiarity for content words and AWL
	13. Correct answer position
	14. Degree of lexical overlap between correct answer and distractors (92%)
Item-by-Text	15. Number of texts needed to answer item (100%)
	16. Item reference (to whole text or part of text) (100%)
	17. Subskill tested (global, local, or inferential) (91%)
	18. Explicitness of requested information (87%)
	19. Location of requested information in text (81%)
	20. Percentage of relevant text to answer question (76%)
	21. Number of plausible distractors (78%)
	22. Level of abstractness of question (79%)

* Percentages between parentheses refer to inter-coder agreement for manually coded variables.

to the extent to which increasingly large amounts of information are incorporated into increasingly short grammatical units (Lu, 2011). *Coh-Metrix* was used to estimate two measures of syntactic complexity for each text: Sentence length (i.e., average number of words per sentence) and syntactic similarity. As Ozuru et al. (2008) explained, sentence length affects processing demand; processing a longer sentence places larger demands on working memory, potentially rendering the text more difficult (cf. Rupp et al., 2001). Syntactic similarity measures the uniformity and consistency of the syntactic constructions in the text (Graesser et al., 2004). One index of syntactic similarity was used: syntactic similarity all, which measures syntactic similarity across all sentences and paragraphs in the text. Generally, high syntactic similarity indices indicate less complex syntax that is easier to process (Crossley et al., 2008; Graesser et al., 2004).

Four groups of measures of the text lexical characteristics were examined: lexical density, lexical variation, lexical sophistication, and word information. *Lexical density* refers to the proportion of content words (i.e., nouns, verbs, adjectives, and adverbs) in a text. A lexical density score was computed for each text. *Lexical variation* refers to the variety of words in a text and is often measured using the Type-Token Ratio (TTR), that is, the ratio of the types (the number of different words used) to the tokens (the total number of words used) in a text (Laufer & Nation, 1995; Malvern & Richards, 2002). A low ratio indicates that the text makes repeated use of a smaller number of types (words), whereas a high TTR suggests that the text includes a large proportion of different words, which can make the text more demanding. One problem with TTRs is that they tend to be affected by text length, which makes them unsuitable measures when there is much variability in text length (Koizumi, 2012; Malvern & Richards, 2002; McCarthy & Jarvis, 2010). To address this issue, the Measure of Textual and Lexical Diversity (MTLD) as computed by *Coh-Metrix* was used. MTLD values do not vary as a function of text length (Koizumi, 2012; McCarthy & Jarvis, 2010), thus, allowing for comparisons between texts of considerably different lengths like the ones in this study.

Lexical sophistication concerns the proportion of relatively unusual, advanced, or low-frequency words to frequent words used in a text (Laufer & Nation, 1995; Meara & Bell, 2001). Two measures were used to assess lexical sophistication: lambda as computed by the *P-Lex* program (Meara & Bell, 2001) and average word length (AWL, average number of characters per word) as computed by *Coh-Metrix*. A low value of lambda shows that the text contains mostly high-frequency words, whereas a higher value indicates more sophisticated vocabulary use (Read, 2005). Higher AWL values indicate more sophisticated vocabulary use (Read, 2005). Finally, three measures of the characteristics of content words used in the texts were obtained from *Coh-Metrix*: word frequency, word familiarity, and word polysemy (Crossley et al., 2008; Graesser et al., 2004; McNamara, Crossley, & McCarthy, 2010). *Word frequency*, measured using the mean CELEX word frequency score for content words,[3] refers to how often particular words occur in the English language (Graesser et al., 2004; Ozuru et al., 2008). As Crossley et al. (2008) explained, frequent words are normally read more rapidly and understood better than infrequent words, which can enhance L2 reading performance (cf. Ozuru et al., 2008). *Word familiarity* refers to how familiar a word is based on familiarity ratings of words by Toglia and Battig (1978) and Gilhooly and Logie (1980). Generally, words that are more familiar are recognized more quickly and sentences with more familiar words are processed faster (Crossley et al., 2008). When a text has a low familiarity score and many infrequent words, readers may experience difficulty understanding the text, resulting in an increased difficulty of the questions associated with the text (Graesser et al., 2004; Ozuru et al., 2008). *Polysemy* is measured as the number of senses a word has (but not which sense of a word is used) using the WordNet computational, lexical database developed by Fellbaum (1998) (Crossley et al., 2007). *Coh-Metrix* reports the mean WordNet polysemy value for all content words in a text.

The study included three indicators of text coherence and cohesion as computed by *Coh-Metrix*: referential cohesion, conceptual cohesion, and connectives density. These indices are based on the assumption, put forward by Graesser et al. (2004), that cohesion is a property of a text that involves "explicit features, words, phrases or sentences that guide the reader in interpreting the substantive ideas in the text, in connecting ideas with other ideas and in connecting ideas to higher level global units (e.g. topics and themes)" (p. 193; cf. Green et al., 2010). *Referential cohesion* refers to the extent to which words in the text co-refer. These types of cohesive links have been shown to aid in text comprehension and reading speed (Ozuru et al., 2008). *Coh-Metrix* provides several measures of referential cohesion, including several measures of argument overlap and content word overlap. Argument overlap measures how often two sentences share common arguments (nouns, pronouns, or noun phrases). As Ozuru et al. (2008) explained, less argument overlap between adjacent sentences places demands on the reader because the reader needs to infer the relations between the sentences to construct a global representation of the text. Content word overlap is the proportion of content words in the text that appear in adjacent sentences sharing common content words. Overlapping vocabulary has been found to be an important aspect in reading processing and can lead to gains in text comprehension and reading speed (Ozuru et al., 2008). Four *Coh-Metrix* measures were examined, argument overlap between adjacent sentences, argument overlap between all sentences in a text, content word overlap between adjacent sentences, and content word overlap between all sentences in a text. The inter-correlations among the four measures were high (0.70 to 0.90). Consequently, only one measure of referential cohesion, argument overlap between adjacent sentences, is included in the study.

Conceptual cohesion concerns the extent to which the content of different parts of a text (e.g., sentences, paragraphs) is similar semantically or conceptually. Text cohesion (and sometimes coherence) is assumed

[3] The CELEX frequency score is based on the database from the Center of Lexical Information (CELEX) which consists of frequencies taken from the early 1991 version of the COBUILD corpus of 17.9 million words (see Crossley et al., 2007, 2008).

to increase as a function of higher conceptual similarity between text constituents (Crossley et al., 2008; Graesser et al., 2004; McNamara et al., 2007). The main measures of this variable are based on Latent Semantic Analysis (LSA). LSA is a statistical, corpus-based technique that provides an index of local and global conceptual cohesion and coherence between parts of a text by considering similarity in meaning, or conceptual relatedness, between parts of a text (i.e., sentences, paragraphs) (Crossley et al., 2008; Foltz, Kintsch, & Landauer, 1998; Graesser et al., 2004; McNamara, Cai, & Louwerse, 2007). Cohesion is expected to increase as a function of higher LSA scores. *Coh-Metrix* computes several LSA measures for each text. Two measures were examined in this study: LSA mean sentence adjacent similarity, which represents the similarity of concepts between adjacent sentences in a text, and LSA mean all sentences similarity, which represents the similarity of concepts between all sentences across a text as a whole (Crossley et al., 2007, 2008). The correlation between the two measures for the 54 texts in this study was 0.82. Consequently, only one measure, LSA mean all sentences similarity, was included.

The last indicator of text coherence and cohesion is *connectives density*. Connectives provide explicit cues to the types of relationships between ideas in a text, thus, providing important information about a text's cohesion and organization (Crossley et al., 2008; Graesser et al., 2004; Graesser, McNamara, & Kulikowich, 2011). For each text, *Coh-Metrix* provides an incidence score (per 1000 words) of all connectives (Crossley et al., 2008; Graesser et al., 2004, 2011).

Variable 9 in Table 3 concerns text concreteness/abstractness. *Coh-Metrix* provides a number of indices that relate to the degree of abstractness of a text based on a detailed analysis of content words in terms of their concreteness (how concrete or abstract a word is), meaningfulness (how many associations a word has with other words), and imageability (how easy it is to construct a mental image of a word in one's mind) (Graesser et al., 2004, 2011). *Coh-Metrix* computes scores for each of these three features for each text, based on a dataset that involves human ratings of thousands of words along several psychological dimensions (Graesser et al., 2004, 2011). Based on these three measures, *Coh-Metrix* then computes a *z*-score for word concreteness for each text. Higher *z*-scores on this dimension indicate that a higher percentage of content words in the text are concrete and meaningful and evoke mental images, as opposed to being abstract (Graesser et al., 2011). Higher *z*-scores thus reflect easier processing (Graesser et al., 2004, 2011). A concreteness *z*-score was computed for each MET text in the study.

Finally, *Flesch Reading Ease* was used to measure text readability. *Coh-Metrix* calculates the Flesch Reading Ease using a formula, reported by Flesch (1948), based on the number of words per sentence and the number of syllables per word (Crossley, Allen, & McNamara, 2011).[4] Flesch Reading Ease scores range from 0 to 100, with lower scores reflecting more challenging texts (Graesser et al., 2004, 2011; Green et al., 2010).

Item and Item-by-Text Variables

As Table 3 shows, the study included four item variables and eight item-by-text variables. Variables 11 and 12 concern item length (total number of words in stem and all options) and item vocabulary. Two measures of the vocabulary of the stem and options were computed using *Coh-Metrix*: word familiarity for content words and average word length (see definitions above). Variable 13 concerns the ordinal position of the correct option (coded as 1, 2, 3, or 4). The last item variable (variable 14) concerns the degree of lexical overlap between the correct option and the distractors for any given item. It was measured by computing the proportion of words in the correct option that overlap with words in the three distractors (cf. Freedle & Kostin, 1993). Items with a higher degree of lexical overlap between correct answer and distractors tend to be harder than items that have lower overlap (Rupp et al., 2001).

The last set of eight variables in Table 3 (variables 15 to 22) concerns the relationships between items and their texts. Variable 15 is a dichotomous variable and concerns the number of texts that the item refers to (one text or multiple texts). Variable 16, *item reference*, concerns whether the item asks the test-taker to refer to the whole text (coded 0) or to a specific section of the text (e.g., in paragraph 3 of text B, the first sentence of paragraph 2; coded 1). Variable 17 concerns the *subskill tested* by each item. Specifically, each item was coded as focusing primarily on one of three subskills as defined by CaMLA (2012): global (understanding main idea, identifying speaker's/author's purpose; synthesizing ideas from different parts of the text); local (identifying supporting detail; understanding vocabulary;

[4] The formula is as follows:
Flesch Reading Ease = 206.835 – (1.015 x number of words/number of sentences) – (84.600 x number of syllables/number of words) (Crossley et al., 2011, p. 90).

synthesizing details; recognizing restatement); or inferential (understanding rhetorical function; making an inference; inferring supporting detail; understanding pragmatic implications). Variable 18, *explicitness of requested information*, asks whether the information needed to answer the question correctly is explicitly (coded 0) or implicitly (coded 1) mentioned in the text (Rupp et al., 2001). Implicit information could be text based or it could be textually relevant background knowledge. Rupp et al. (2001) noted that inferencing is more cognitively demanding than recognizing explicitly stated information.

Location of requested information in text (variable 19) refers to the location in the text of information relevant to the question (Gorin & Embreston, 2006; Rupp et al., 2001). Each text was divided into three equal parts (based on word count). Next, the section of the last occurrence of the correct information in the text is coded as early, middle, late, entire text, or multiple texts (cf. Rupp et al., 2001). As Rupp et al. (2001) noted items are more difficult if information is located earlier in the text, because the information may no longer be in one's short-term memory. *Percentage of relevant text* (variable 20) refers to the proportion of the text necessary for correctly responding to the question (Gorin & Embreston, 2006). To code this variable, the relevant portion of the text needed to correctly answer a question was identified. Next, the number of words in the relevant portion was counted and divided by the total number of words in the text. Items requiring information from the entire text were scored as 100%.

Number of plausible distractors (variable 21) concerns the number of distractors (out of 3) that contain ideas that are either directly addressed in the text or that can be inferred from the text. Distractors that include words and/or propositions that overlap with words and/or propositions in the text are coded as plausible. The number of these plausible distractors (0 to 3) was then counted for each item (Ozuru et al., 2008; Rupp et al., 2001). As Rupp et al. (2001) noted, items are more difficult if the number of plausible distractors increases since finer distinctions will be needed to identify the requested information. Finally, variable 22, *level of abstractness*, measures the level of abstractness or concreteness of the information requested by the question (Ozuru et al., 2008; Rupp et al., 2001). Ozuru et al. (2008) argued that searching for abstract, as opposed to concrete, information in a text tends to require a more extensive search and more information integration, rendering the task more difficult (cf. Rupp et al., 2001). Each item was assigned to one of five levels of abstractness: (0) *Most concrete* questions ask for the "identification of persons, animals, or things;" (1) *Highly concrete* questions ask for the "identification of amounts, times, or attributes;" (2) *Intermediate* questions ask for the "identification of manner, goal, purpose, alternative, attempt, or condition;" (3) *Highly abstract* questions ask for the "identification of cause, effect, reason, result" or evidence; or (4) *Most abstract* questions ask for the "identification of equivalence, difference, or theme" (from Mosenthal, 1996, cited in Ozuru et al., 2008, p. 1004).

Statistical Analyses

To address research question 1, descriptive statistics for the text and item variables in Table 3 were computed. Additionally, because the texts and items in the MET reading subsection are organized by section (or text type), texts and items in the three sections (A, B, and C) were compared in terms of the various continuous measures in Table 3 using analysis of variance (ANOVA), with text section as the independent variable and each of the text and item measures as a dependent variable. Where a significant difference was detected across sections, follow-up pairwise comparisons (using a Bonferroni correction) were conducted. Furthermore, in an attempt to better understand the meaning of some of the *Coh-Metrix* indices concerning text characteristics, some of these indices are compared to findings from three previous studies that used *Coh-Metrix* to analyses reading texts for ESL learners (Crossley et al., 2007; Crossley & McNamara, 2008; Green et al., 2010). Crossley et al. (2007) compared the linguistic and discourse characteristics of 81 simplified texts from 7 beginning ESL grammar, reading, and writing textbooks and 24 authentic texts from 9 textbooks. Crossley and McNamara (2008) compared 123 simplified texts and 101 authentic texts from 11 intermediate ESL and EFL reading textbooks for adult ESL/EFL learners. Green et al. (2010) analyzed and compared 42 texts from 14 core undergraduate textbooks at one university in the U.K. and 42 texts from 14 IELTS Academic reading tests in terms of various features that affect text difficulty in order to evaluate the authenticity of IELTS reading texts.

Second, to address research questions 2 and 4, item scores from the 6 MET forms and 6,250 test takers in the study were analyzed using the computer program *FACETS* (Linacre, 2011), which operationalizes the Multi-faceted Rasch Model, in order to (a) estimate item difficulty and item fit and (b) identify and estimate any significantly biased interactions between reading items in each form and test-taker subgroups, defined in terms of L1 (Spanish, Albanian, or Other), age range (11–18, 19–24, 25–30, 31–40, or over 40), and gender. Items were then grouped in terms of whether they showed fit or misfit and whether they were involved in biased interactions or not. Fit values within two standard deviations from the mean were considered to represent adequate fit. Because the six different forms of the test in the study were administered to six different groups of test takers, the dataset included six unlinked datasets, one for each form/group. In order to provide enough connection in the dataset for the *FACETS* analysis to run, it was necessary to link the six datasets to allow comparisons across forms (Linacre, 2011). This was done by assuming that, since the forms were developed based on the same test specifications, the six forms were equivalent in terms of their difficulty.

Most previous studies on the relationships between item and text features and item difficulty (i.e., research question 3 in this study) used multiple regression analysis. However, this technique ignores the nested structure of reading test data. Nested data means that observations at lower levels are nested within units at higher levels (Hox, 2002). The lowest level of the hierarchy is labeled level 1, the next level, level 2, and so forth (Hox, 2002). Reading test data often have a 2-level nested structure with items being nested within texts (Ozuru et al., 2008). Ignoring the nested structure of reading test data violates the assumption of independence of observations underlying multiple regression analysis, which can lead to biased estimates of relationships among variables and to Type I error. Multilevel modeling (MLM) addresses this problem (Barkaoui, 2013b; Hox, 2002; Luke, 2004).

MLM views reading item data as nested within text. It distinguishes between two levels of analysis: level-1 observations (items) nested in level-2 units (texts). Given an outcome variable, such as item difficulty, a level-1 equation examines how the outcome relates to item-level factors (e.g., item length). The relationship between item-level factors and item difficulty within text can vary across texts. At level 2, text-level factors (e.g., text length) serve as predictor variables. MLM, thus, provides increased power in the prediction of outcomes and correct estimates of relationships among variables at multiple levels. It also allows the simultaneous examination and separation of the main and interaction effects of variables at different levels of analysis to variability in item difficulty (Barkaoui, 2013b; Hox, 2002; Luke, 2004; Ozuru et al., 2008). Examining the effects of the interactions between level 1 (item) and level 2 (text) variables on item difficulty (e.g., the influence of text features on the contribution of item features to item difficulty) is important because the difficulty level of an item will depend not only on the item itself, but also on the text on which the item is based (Ozuru et al., 2008). Practically, MLM can handle unbalanced designs such as texts with different number of items.

To address research question 3 concerning the relationships between item, text, and item-by-text variables, on one hand, and item difficulty indices, on the other, various statistical analyses were conducted. First, Pearson *r* correlations between continuous text and item variables, on the one hand, and indices of item difficulty, on the other, were examined. Second, item difficulty estimates were statistically compared across categories of categorical item and text variables (using ANOVA). When ANOVA results were significant, follow-up pairwise comparisons using a Bonferroni correction were conducted. Only item and text variables that were found to be significantly associated with item difficulty (based on ANOVA and correlational analyses) were included in MLM analyses in order to estimate the contribution of each item and text feature to item difficulty. For all MLM analyses, the computer program *HLM6* (Raudenbush, Bryk, Cheong & Congdon, 2004) was used. The data in this study were considered to consist of two levels.[5] As will be described below, five sequential models were estimated following Hox's (2002) recommendations. For each model, two main indices were examined: the deviance statistic which compares the fit of multiple models to the same dataset and significance tests for individual coefficients (Hox, 2002; Luke, 2004). Based on the results of these different models, a final model was built.

A similar process was adopted to address research question 5. Specifically, analysis of variance (ANOVA), cross-tabulation, chi-square tests, and logistic MLM were used to identify the item, text, and item-by-text features that characterize items involved in significantly biased interactions with test-taker subgroups. Logistic MLM is used when the outcome variable is binary (e.g., whether

[5] Strictly speaking, the dataset in this study has a 4-level structure, item nested within text nested within reading set nested within form. However, preliminary analyses indicated that there were no significant differences across sets and forms in terms of item difficulty. To simplify the analyses, only two levels, items and text, are considered in this study.

an item exhibits significantly biased interaction with a facet, coded 1, or not, coded 0) and works by including a logit transformation of the outcome variable (and appropriate error distribution) in order to estimate the contribution of the predictors to the probability or likelihood that the outcome is 1 (e.g., item showing significantly biased interaction with test taker L1) (Hox, 2002; Luke, 2004).

Findings

Description of the Linguistic and Discourse Characteristics of MET Reading Texts and Items

This section reports and discusses descriptive statistics concerning the linguistic and discourse characteristics of the sample of MET texts and items included in the study.

MET Text Characteristics

There was an equal number of texts for each text type or section (*n* = 18, 33%), but most of the 54 texts were related to the public (*n* = 34, 63%) and occupational (*n* = 18, 33%) domains. About two thirds of these texts (*n* = 33, 61%) included nonverbal information. The texts covered five topics: economic and job-related (*n* = 16, 30%), everyday life (*n* = 14, 26%), health/psychology (*n* = 8 or 15%), environment (*n* = 8 or 15%), and science and technology (*n* = 8 or 15%). Table 4 displays descriptive statistics for the continuous text variables in the study for all the texts as well as for each section.

The MET texts included in this study varied in terms of length between 49 and 311 words (*M* = 183 words). As noted above, by design, section A texts are the shortest (*M* = 87.22 words) and section C texts are the longest (*M* = 289.94). There was some variability in terms of syntactic complexity across texts, particularly in terms of sentence length which varied significantly across sections (F[2, 51]= 38.24, $p < 0.05$, $\eta^2 = 0.60$).[6] Follow-up pairwise comparisons (using a Bonferroni correction) indicated that sentences in section C texts were significantly longer (*M* = 17.45 words) than those in section B texts (*M* = 13.25), which were in turn significantly longer than those in section A texts (*M* = 8.91). It should be noted here that there was a high and significant correlation between text length (number of words) and sentence length (average number of words per sentence) as estimated by *Coh-Metrix* ($r = 0.79$). The differences in sentence length seem to be due mainly to differences in text length across sections. One way to clarify the meaning of the sentence length indices is to compare the sentence length indices in Table 4 to those reported in other studies using *Coh-Metrix*. For example, the texts in this study have much shorter sentences (*M* = 13.21 words per sentence) than those in the authentic texts (*M* = 21.47) and IELTS reading texts (*M* = 21.89) in Green et al. (2010). Finally, as Table 4 shows, differences across sections in terms of the second measure of syntactic complexity, syntactic similarity, were not significant. As noted above, syntactic similarity measures the uniformity and consistency of the syntactic constructions in the text (Graesser et al., 2004).

None of the measures of text lexical features examined in this study (i.e., lexical density, variation, sophistication, and word information) varied significantly across sections. In order to better interpret the lexical indices in Table 4, it may be helpful to compare them to those reported in other studies using *Coh-Metrix* while bearing in mind the differences across the contexts and purposes of those studies. For instance, the lexical density of the texts in this study (*M* = 0.58) is closer to the lexical density of the IELTS reading texts (*M* = 0.57) than to that of the authentic texts (*M* = 0.56) in Green et al. (2010). Text average word length in this study (*M* = 5.00) is also closer to the IELTS reading texts (*M* = 5.03) than to that for the authentic texts (*M* = 5.14) in that study. It seems the MET reading texts are similar to those in IELTS Academic, at least in terms of some aspects of their lexical density and sophistication as measured in this study. Word information measures also did not vary significantly across tasks, but all three indices are lower than those reported by Crossley et al. (2007) and Crossley and McNamara (2008) for authentic and simplified texts for ESL learners, particularly for word polysemy. For example, word familiarity indices in this study (*M* = 568.46) are lower than those for authentic texts (*M* = 576.92) and simplified texts (M = 584.16) in Crossley et al. (2007). Word frequency indices in this study (*M* = 2.08) are also lower than those reported in that study for authentic (*M* = 2.36) and simplified (*M* = 2.44) texts. Finally, word polysemy scores in this study (*M* = 3.79) are much lower than those reported by Crossley et al. (2007) for authentic (*M* = 7.61) and simplified (*M* = 7.35) texts.

[6] *Partial eta-squared* (*partial* η^2) is used as a measure of effect size. *Partial* $\eta^2 \geq 0.01$ indicates a small effect size; *partial* $\eta^2 \geq 0.09$ indicates a medium effect; and *partial* $\eta^2 \geq 0.25$ indicates a large effect (Field, 2009).

Table 4: Descriptive Statistics for Continuous Text Variables for all Texts ($N = 54$) and by Section ($n = 18$ texts per section)

	All Texts				Section A		Section B		Section C	
Variable	M	SD	Min.	Max.	M	SD	M	SD	M	SD
Text length	183.26	85.86	49	311	87.28	20.62	172.56	15.11	289.94	20.11
Syntactic complexity										
Sentence length*	13.21	4.54	5.44	23.00	8.91	2.04	13.25	3.23	17.45	3.34
Syntactic similarity	0.09	0.02	0.05	0.15	0.10	0.03	0.10	0.03	0.08	0.02
Lexical Features										
Density	0.58	0.05	0.46	0.78	0.59	0.08	0.57	0.06	0.58	0.04
Variation	103.81	28.28	21.98	176.61	104.53	34.42	103.83	25.05	103.07	26.08
Sophistication										
lambda	1.46	0.48	0.44	2.45	1.32	0.47	1.44	0.53	1.62	0.42
AWL	5.00	0.46	3.92	6.51	4.92	0.59	4.97	0.49	5.12	0.26
Word information										
Frequency	2.08	0.17	1.52	2.40	2.02	0.21	2.10	0.16	2.13	0.12
Familiarity	568.46	8.73	542.97	585.95	570.97	11.33	568.21	7.82	566.20	6.02
Polysemy	3.79	0.46	2.56	4.87	3.59	0.48	3.85	0.48	3.93	0.38
Coherence/Cohesion										
Referential*	0.35	0.23	0.00	0.92	0.16	0.12	0.40	0.22	0.49	0.19
Conceptual*	0.20	0.13	0.00	0.59	0.14	0.14	0.22	0.13	0.24	0.12
Connectives density*	80.77	22.36	34.48	147.06	69.79	23.59	87.61	23.78	84.91	15.56
Text concreteness	0.35	0.95	-1.33	2.47	0.67	1.07	0.20	0.95	0.19	0.79
Readability										
Flesch Reading Ease*	54.86	12.62	18.50	81.38	60.48	14.51	56.06	12.39	48.03	6.95

* Variable showed statistically significant differences across sections.

Overall, the texts in this study included more words that are less familiar, less frequently used, and have fewer senses (i.e., lower polysemy values) than the authentic and simplified texts for beginner and intermediate ESL learners examined by Crossley et al. (2007) and Crossley and McNamara (2008).

All three measures of text coherence and cohesion varied significantly across sections: Referential cohesion (F[2, 51]= 15.56, $p < 0.05$, $\acute{\eta}^2 = 0.38$), conceptual cohesion (F[2, 51] = 3.29, $p < 0.05$, $\acute{\eta}^2 = 0.11$), and connectives density (F[2, 51] = 3.65, $p < 0.05$, $\acute{\eta}^2 = 0.13$). For referential cohesion, section A texts had a significantly lower argument overlap between adjacent sentences ($M = 0.16$) than section B ($M = 0.40$) and section C ($M = 0.49$) texts. The difference between section B and section C texts was not significant. Concerning conceptual cohesion, measured using LSA mean all sentences similarity, there was a significant difference between section A texts ($M = 0.14$) and section C texts ($M = 0.24$). As for connectives, there was a significant difference between section A texts ($M = 69.79$ connectives per 1000 words) and section B texts ($M = 87.91$). By way of comparison, the conceptual cohesion of the texts in this study ($M = 0.20$) seems similar to that for IELTS reading texts ($M = 0.21$) in Green et al. (2010) and the simplified ESL reading texts

(M = 0.18 to 0.20) in Crossley et al. (2007) and Crossley and McNamara (2008), but is lower than the conceptual cohesion of the authentic texts (M = 0.26) in Green et al. (2010) and higher than those reported in the other two studies (M = 0.16). The connective density of the MET texts (M = 80.77 connectives per 1000 words) was much higher than those for authentic and simplified beginner ESL texts (M = 71 to 73) in Crossley et al. (2007), but similar to those for intermediate ESL texts (M = 80 to 82) reported by Crossley and McNamara (2008). These patterns suggest that the three sections differ in terms of their cohesion and coherence, perhaps as a function of variation in text length and type. The correlations between text length, on the one hand, and referential cohesion and conceptual cohesion, on the other, were 0.60 and 0.37, respectively, while the correlation between referential cohesion and conceptual cohesion was 0.47. Additionally, the MET texts in this study seem to be similar to IELTS Academic texts and simplified ESL reading texts in terms of their conceptual cohesion. The MET texts also seem to be similar to intermediate ESL texts in terms of the number of connectives. Finally, there were no significant differences across sections in terms of text concreteness/abstractness.

There was a significant difference across sections in terms of Flesch Reading Ease (F[2, 51]= 5.22, p < 0.05, $\dot{\eta}^2$ = 0.17). As Table 4 shows, section C texts have a lower index than section B texts which have a lower index than section A texts. However, only the difference between section A texts (M = 60.48) and section C texts (M = 48.03) was statistically significant. This suggests that section C texts are significantly more challenging than section A texts.[7] One way to clarify the meaning of the readability indices in Table 4 is to compare them to those reported in other studies using *Coh-Metrix*. Thus, the texts in this study seem to be much easier (Flesch Reading Ease M = 54.86) than the authentic texts (M = 36.82) and IELTS reading texts (M = 42.15) in Green et al. (2010). This is not surprising given that IELTS targets higher proficiency levels (levels B1 to C2 on CEFR) compared to MET (levels A2 to C1).

MET Item and Item-by-Text Characteristics

Of the 216 reading items included in the study, 192 (89%) related to single texts, while 24 (11%) asked about multiple texts. The majority of the items tested local understanding (n = 124, 57%), followed by inferencing (n = 48, 22%) and global understanding (n = 44, 20%). The majority of the items (n = 164, 76%) referred to the whole text; about a quarter of the items (n = 52, 24%) referred to a specific part of a text. There were 114 (53%) items that required information that is explicitly mentioned in the text; the remaining 102 (47%) items were about implicit information. In terms of the ordinal position of the correct answer, this was distributed fairly equally across positions with 52 (24%) items in the first position, 65 (30%) in the second position, 42 (19%) in the third position, and 57 (26%) in the fourth position. The location of required information in the text tended to be mainly in the middle (n = 53, 25%), last third (n = 54, 25%), or throughout the text (n = 50, 23%). About a sixth of the items (n = 35, 16%) related to information in the first third of the text and the remaining 24 items (11%) asked about multiple texts.

Table 5 displays descriptive statistics for the continuous item and item-by-text variables in the study for all the items as well as for each section. ANOVA was used to compare items associated with the three sections (A, B and C) and multiple texts in terms of the measures of the item and item-by-text characteristics listed in Table 5. There was a significant difference across sections in terms of item length (F[3, 212]= 6.42, p < 0.05, $\dot{\eta}^2$ = 0.08). As Table 5 shows, items associated with section A texts are significantly shorter than those associated with section B and with multiple texts. Section C items were also significantly shorter than those associated with multiple texts. There was no significant difference in item length (i.e., number of words per item) between sections B and C.

Overall, question AWL (M = 4.92) and word familiarity (M = 569.89) are very similar to those for the texts (M = 5.00 and 568.46, respectively). However, question word familiarity (F[3, 212] = 2.81, p < 0.05, $\dot{\eta}^2$ = 0.04) and average word length (F[3, 212] = 7.30, p < 0.05, $\dot{\eta}^2$ = 0.09) varied significantly across sections C and B, with items in section B having significantly higher word familiarity scores but significantly lower

[7] It should be noted here that Flesch Reading Ease is moderately correlated with text length (r = 0.40).

Table 5: Descriptive Statistics for Continuous Item and Item-by-Text Variables for all Items and across Text Types (Sections)

	Total (N = 216)				**Section A** (n = 42)		**Section B** (n = 67)		**Section C** (n = 83)		**Multiple Texts** (n = 24)	
Variable	M	SD	Min.	Max.	M	SD	M	SD	M	SD	M	SD
Item Variables												
Item length*	32.53	9.56	11	59	28.31	8.76	33.10	9.30	32.47	9.77	38.54	7.64
Item vocabulary												
Word familiarity*	569.89	15.48	515.07	602.75	572.06	15.20	572.62	14.22	566.09	15.66	571.68	16.99
Average word length*	4.92	0.56	3.52	6.54	4.88	0.69	4.72	0.49	5.13	0.52	4.85	0.41
Lexical overlap of correct answer & distractors	0.11	0.16	0.00	0.75	0.09	0.14	0.08	0.15	0.15	0.18	0.12	0.14
Item-by-Text Variables												
Level of abstractness*	2.12	1.18	0	4	1.79	1.05	1.79	1.21	2.20	1.09	3.33	0.70
Percentage of relevant text*	42.64	40.39	2.74	100	49.08	40.04	33.85	35.70	29.90	35.21	100.00	0.00
Number of plausible distractors	1.50	0.88	0	3	1.38	0.91	1.51	0.89	1.51	0.89	1.62	0.82

* Variable showed statistically significant differences across sections.

AWL. The level of abstractness of information required by the question also varied significantly across sections (F[3, 212] = 13.48, $p < 0.05$, $\dot{\eta}^{2} = 0.16$). In particular, items about multiple texts tended to require more abstract information than items about single texts in each section. There was also a significant difference in terms of the percentage of text needed to answer the question across sections (F[3, 212] = 27.92, $p < 0.05$, $\dot{\eta}^{2} = 0.28$). By design, items based on multiple texts required reading all the texts (M = 100%), while the other items, on average, required reading between 30% (section C) and 49% (section A) of the text. Consequently, significant differences were detected between multiple-text items and items in each section as well as between section C and section A items. This is not surprising given that multiple-text items tend to be more inferential and global in nature. It is worth noting here that all the items about multiple texts (n = 24) and most of the items on section A texts (n = 38 out of 42) referred to the whole text. About a third of the items in sections B and C, in contrast, explicitly referred the test-taker to specific parts of the texts. Both the degree of lexical overlap between the correct answer and distractors and the number of plausible distractors did not vary significantly ($p > 0.05$) across sections.

Finally, items assessing different reading subskills were compared across sections and in relation to other item variables. The distribution of the 216 items in terms of subskills tested across sections in the 6 test forms included in this study was as follows: Section A included 12 (29%) global, 23 (55%) local, and 7 (17%) inferential items. Section B included 15 (22%) global, 45 (67%) local, and 7 (10%) inferential items. Section C included 17 (21%) global, 56 (68%) local, and 10 (12%) inferential items. Comparisons of the seven continuous item variables across subskills tested (i.e., global, local or inferential) indicated that items addressing different subskills differed significantly in terms of their length (F[2, 213] = 80.26, $p < 0.05$, $\acute{\eta}^2 = 0.13$), lexical overlap between the correct answer and the distractors (F[2, 213] = 10.52, $p < 0.05$, $\acute{\eta}^2 = 0.09$), the percentage of text needed to answer the question (F[2, 213]= 222.01, $p < 0.05$, $\acute{\eta}^2 = 0.68$), and the level of abstractness of the information required by the question (F[2, 213] = 44.43, $p < 0.05$, $\acute{\eta}^2 = 0.29$). First, inferential items were significantly longer (*M* = 38.4 words, *SD* = 7.38) than global (*M* = 33.59, *SD* = 7.10) and local (*M* = 29.89, *SD* = 10.03) items. The local items, however, had a larger variability in terms of their length. Second, global items had significantly greater lexical overlap between the correct answer and the distractors (*M* = 0.20, *SD* = 0.17) than the local (*M* = 0.08, *SD* = 0.15) and inferential (*M* = 0.14, *SD* = 0.16) items. Third, global items required reading a significantly greater portion of the text (*M* = 91%, *SD* = 25) than did the inferential items (*M* = 71%, *SD* = 39), which in turn required reading a significantly greater portion of the text than did the local items (*M* = 15%, *SD* =11). Fourth, not surprisingly, local items had a significantly lower level of abstractness (*M* = 1.57, *SD* = 1.00) than did inferential (*M* = 2.94, *SD* = 0.98) and global (*M* = 2.77, *SD* = 1.01) items. There were no significant differences across itesm testing different subskills in terms of question word familiarity, question AWL, and the number of plausible distractors.

MET Item Difficulty, Fit, and Bias Indices

This section reports and discusses findings concerning the difficulty, fit, and bias indices of the sample of 216 MET items included in the study to address research questions 2 and 4. Table 6 summarizes *FACETS* results concerning measures and fit statistics for test-taker ability and item difficulty. It shows that the test takers varied in terms of their ability estimates between -2.61 and 5.38 logits (*M* = 0.49, *SD* = 1.31). The positive mean ability estimate suggests that the test was slightly easy for this group of test takers. The X^2 test, which tests the hypothesis that all test takers are equal in terms of the ability being measured, is statistically significant at $p < 0.001$. The strata and reliability indices for the difference in test-taker ability are high (3.81 and 0.87, respectively), indicating that the variance among test takers is substantially larger than the error of estimates and that the test separates the test takers into approximately four statistically distinct levels in terms of the ability being measured. The high reliability

Table 6: Summary of FACETS Measures and Fit for Test-taker Ability and Item Difficulty

	Test-Taker Ability	**Item Difficulty**
Sample	6,250 test takers	216 items
Estimates (in logits)		
M (Model SE)	0.49 (0.43)	0.00 (0.08)
SD (Model SE)	1.31 (0.18)	0.75 (0.01)
Minimum	-2.61	-2.05
Maximum	5.38	2.78
Range	7.99	4.83
Infit M (SD)	1.00 (0.12)	1.00 (0.10)
Outfit M (SD)	1.00 (0.26)	1.00 (0.19)
Separation Statistics		
Separation	2.61	9.82
Strata	3.81	13.42
Reliability of Separation	0.87	0.99
Fixed Chi-Square Statistic	39267.7	18330.5
df.	6245	215
Significance	0.00	0.00

statistic indicates that the same ordering of test takers would be likely to be obtained if test takers were to take another test measuring the same ability. High test-taker strata and reliability indices mean that the assessment distinguishes between test takers in terms of the ability being measured and that one can place confidence in the replicability of test-taker placement across other tasks or tests that measure the same construct (Bond & Fox, 2007). This means greater confidence in the consistency of score-based inferences.

Table 6 shows that the 216 items spanned almost 5 logits in terms of difficulty (i.e., -2.05 to 2.78 logits). Because item difficulty was centered during *FACETS* analyses, the mean difficulty of the items is 0.00 (*SD* = 0.08). The items differed significantly in terms of their difficulty as indicated by the high reliability (0.99) and strata (13.42) indices and the significant X^2 statistic ($p < 0.01$). The analysis reliably separated the items into more than 13 levels of difficulty. The item reliability index indicates the replicability of item placements relative to each other in terms of difficulty if these same items were given to another sample with comparable ability levels. The high reliability indicates that the analysis is reliably separating items into different levels of difficulty (Bond & Fox, 2007). Fit statistics for items ranged between 0.81 and 1.31 for infit MNSQ (*M* = 1.0, *SD* = 0.10) and between 0.63 and 1.70 for outfit MNSQ (*M* = 1.0, *SD* = 0.19). Using the criterion of fit values within two standard deviations from the mean for adequate fit (i.e., 0.80 to 1.20 for infit MNSQ and 0.62 to 1.38 for outfit MNSQ), only two (2) items showed misfit (i.e., infit MNSQ above 1.20); all other items (*n* = 214) had fit statistics within the acceptable range.

Table 7 reports the results for bias analyses. As recommended by McNamara (1996) and Kondo-Brown (2002), only biased interactions (a) with *z*-values equal to or higher than the absolute value of 2 and (b) with infit mean square (infit MNSQ) values within the range of two standard deviations around the mean of infit are considered to be significant. Infit MNSQ shows how consistent the pattern of bias is across all test takers involved in a biased interaction. Table 7 shows that there were no significantly biased item-by-test-taker interactions and only three biased item-by-gender interactions (involving 2 items). There were 96 significantly biased item-by-age group interactions (8.88% of all interactions) and 107 significantly biased item-by-L1 interactions (16.5% of all interactions). Because the number of items with misfit and the number of items involved in biased interactions with test-taker gender were small, analyses in the following section focus only on the relationships between item and text characteristics, on one hand, and item difficulty measures and biased item-by-age and item-by-L1 interactions, on the other.

Table 8 displays the distribution of biased interactions by *z*-value and age and L1 groups. A *z*-value below -2 indicates that the item was significantly easier

Table 7: Summary of Results of Bias Analyses

Biased Interactions	**Item-by-Gender**	**Item-by-Age Group**	**Item-by-L1**	**Item-by-Test-taker**
Number of interactions	432	1080	648	221,486
Number of significantly biased interactions	3	96	107	0
Percentage of biased interactions	0.7%	8.88%	16.5%	0%
Number of items involved in biased interactions (%)	2 (1%)	73 (33.8%)	91 (42%)	0 (0%)
Bias Size M (SD)	0.00 (0.23)	0.00 (0.45)	-0.04 (0.65)	0.03 (0.62)
Model SE M (SD)	0.54 (1.15)	0.32 (0.25)	0.63 (0.72)	2.61 (1.16)
Infit M (SD)	0.90 (0.30)	1.00 (0.20)	0.90 (0.30)	0.50 (0.30)
Outfit M (SD)	0.90 (0.30)	1.00 (0.30)	0.90 (0.40)	0.50 (0.30)
Separation Statistics				
Fixed Chi-Square Statistic	373.2	1705.7	1714.4	20521
df.	504	1296	684	221486
Significance	1.00	0.00	0.00	1.00

Table 8: Number of Significantly Biased Interactions by Z-value and L1 and Age Group

	No Bias	Significantly Biased Interactions		Total
		z-value < -2	z-value > +2	
L1 group				
Albanian	139	42	35	216
Spanish	199	10	7	216
Other L1	203	7	6	216
Total	541	59	48	648
Age group				
11–18 years	177	20	19	216
19–24 years	205	6	5	216
25–30 years	202	6	8	216
31–40 years	202	6	8	216
41 and over	198	6	12	216
Total	984	44	52	1080

for the group, while a *z*-value larger than +2 indicates that the item was significantly more difficult for the group than is normal for that item with all other groups. For L1, there were more biased interactions involving Albanian test takers than the other groups, while for age, there were more biased interactions involving text-takers who are 11 to 18 years old than the other groups. The distribution of negatively (i.e., item was easier) and positively (i.e., item was more difficult) biased interactions within each age and L1 group is not very different. For example, there were 42 negatively biased interactions and 35 positively biased interactions involving Albanian test takers and 20 negatively and 19 positively biased interactions involving test takers who were between 11 and 18 years old. As Table 7 shows, the significantly biased item-by-age group interactions involved 73 items (34% of 216 items), while the significantly biased item-by-L1 interactions involved 91 items (42% of all items). Each item that was involved in one or more biased item-by-age and item-by-L1 interactions was coded 1; all other items were coded 0 for bias.

Relationships between Text and Item Characteristics and Item Difficulty Indices

This section reports and discusses findings concerning the relationships between the linguistic and discourse characteristics of MET texts and items, on the one hand, and item difficulty estimates, on the other, to address research question 3. Table 9 reports the Pearson *r* correlations between continuous item and text variables and item difficulty estimates. It shows that only two text variables, text length ($r = 0.23$) and text connectives density ($r = 0.14$), have significant ($p < 0.05$) correlations with item difficulty estimates. Both correlations are positive indicating that longer texts and texts with more connectives per 1000 words were associated with more difficult items. The correlation between text length and connectives density was $r = 0.23$. Surprisingly, none of the other text variables measured in this study was significantly correlated with item difficulty estimates. Two item variables, question word familiarity ($r = -0.14$) and number of plausible distractors ($r = 0.20$) were significantly ($p < 0.05$) correlated with item difficulty estimates. Generally, items that have higher scores on word familiarity tended to be easier than items with lower familiarity scores, while items with more plausible distractors tended to be more difficult than those that have fewer plausible distractors.

Table 10 displays descriptive statistics (*N*, *M* and *SD*) for item difficulty estimates across categories of categorical text and item variables. In order to identify significant differences in item difficulty across categories, univariate analyses of variance (ANOVA) were conducted with item difficulty as the dependent variable and the categorical variable as the independent variable. Where a significant difference was detected

Table 9: Pearson *r* Correlations of Continuous Text and Item Variables with Item Difficulty Estimates

Text Variables	Item Difficulty	Text Variables	Item Difficulty
Text length	0.23**	Coherence/Cohesion	
Syntactic complexity		Referential	-0.05
Sentence length	0.10	conceptual	0.00
Syntactic similarity	0.05	Connectives density	0.14*
Lexical Features		Concreteness	-0.02
Density	0.04	Readability	
Variation (MTLD)	0.03	Flesch Reading Ease	-0.10
Sophistication		**Item & Item-by-Text Variables**	**Item Difficulty**
lambda	0.03	Item Length	-0.04
AWL	0.07	Item vocabulary	
Word information		Word familiarity	-0.14*
Frequency	0.12	Average word length	0.02
Familiarity	-0.01	Lexical overlap of correct answer & distractors	-0.10
Polysemy	0.07	Level of abstractness	0.00
		Percentage of relevant text	-0.10
		Number of plausible distractors	0.20**

* p < 0.05; ** p < 0.01

for variables with more than two categories, follow-up pairwise comparisons (using a Bonferroni correction) were conducted. These analyses detected significant differences in item difficulty for two text variables: section (F[3, 212] = 4.41, $p < 0.05$, $\hat{\eta}^2 = 0.06$) and nonverbal information (F[1, 214] = 4.71, $p < 0.05$, $\hat{\eta}^2 = 0.02$). Bonferroni follow-up analyses indicated that there was a significant difference between the average difficulty of section A items (*M* = -0.32) and section C items (*M* = 0.16). There were no significant differences between Section B items (*M* = -0.05) and items in sections A or C. Items related to texts that included nonverbal information were significantly more difficult (*M* = 0.70) than items without nonverbal information (*M* = -0.16). It should be noted here that there was a significant association between nonverbal information and text section ($X^2 = 8.88$, *df.* = 2 $p < 0.05$). Specifically, the majority of section C texts included in the study (16 out of 18) contained nonverbal information, while only half section A texts (*n* = 9) and section B texts (*n* = 8) included such information. The significant differences between texts with and without nonverbal information in terms of item difficulty, thus, are most likely due to the characteristics of section C texts (e.g., text length) compared to section A and B texts. Reassuringly, there were no significant differences across text topics or domains in terms of item difficulty.

Only three item variables in Table 10 were associated with significant ($p < 0.05$) differences in terms of item difficulty estimates. First, there was a significant effect of whether the question refers to the whole text or part of the text: (F[1, 214] = 14.17, $p < 0.05$, $\hat{\eta}^2 = 0.06$). Items that referred to part of the text were significantly more difficult (*M* = 0.33) than items that referred to the whole text (*M* = -0.11). Second, subskill tested had a significant effect on item difficulty estimates: F[2, 213] = 3.36, $p < 0.05$, $\hat{\eta}^2 = 0.03$. Bonferroni follow-up analyses indicated that items assessing global understanding (*M* = -0.26) were significantly easier than items assessing local understanding (*M* = 0.07). There were no significant differences between inferential items (*M* = 0.05) and global and local items. The third significant effect concerned whether the information required to answer the question is explicitly or implicitly mentioned in the text (F[1, 214]= 4.42, $p < 0.05$, $\hat{\eta}^2 = 0.02$). Items requesting implicit information were

Table 10: Mean Comparisons for Item Difficulty across Levels of Categorical Text and Item Variables

Text Variable	N	M	SD	Item 7 Item-by-Text Variables	N	M	SD
Domain				Correct answer position	52		
Public	34	0.03	0.72	First	65	0.04	0.75
Occupational	18	-0.07	0.81	Second	42	-0.08	0.63
Educational	2	0.14	0.76	Third	57	-0.02	0.65
Section*				Fourth		0.07	0.93
A	18	-0.32	0.72	Number of texts			
B	18	-0.05	0.78	One text	192	-0.02	0.77
C	18	0.16	0.74	Multiple texts	24	0.12	0.60
Nonverbal information*				Reference*			
Yes	33	0.70	0.72	Whole text	164	-0.11	0.67
No	21	-0.16	0.79	Part of text	52	0.33	0.89
Topic				Subskill tested*			
Health/Psychology	8	-0.17	0.54	Global	44	-0.26	0.76
Environment	8	0.11	0.64	Local	124	0.07	0.81
Economic	16	0.04	0.75	Inferential	48	0.05	0.50
Science/Technology	8	-0.13	0.91	Implicit-Explicit*			
Everyday life	14	0.05	0.81	Explicit	114	-0.10	0.78
				Implicit	102	0.11	0.70
				Location of requested information			
				Early	35	0.21	0.83
				Middle	53	0.02	0.72
				Late	54	-0.03	0.77
				Entire text	50	-0.20	0.73
				Multiple texts	24	0.12	0.60

* indicates statistically significant differences in item difficulty across categories.

significantly more difficult (M = 0.11) than those requiring explicit information (M = -0.10). Number of texts, position of correct answer, and location of required information did not have significant effects on item difficulty.

As noted above, MLM was used to examine the relationships between the linguistic and discourse characteristics of MET texts and items, on one hand, and item difficulty estimates, on the other. Because the number of item and text variables in the study was large, it was necessary to select only a subset of item and text variables for inclusion in MLM analyses. Consequently, only variables that had significant ($p < 0.05$) associations with item difficulty estimates (based on the correlation and ANOVA results above) were selected for inclusion in MLM analyses. The selected variables included four text variables (text length, connectives density, section, and presence of nonverbal information) and five item and item-by-text variables (question word familiarity, item reference to whole or part of text, subskill tested, explicitness of information requested by question, and number of plausible distractors). Next, correlations among the variables in each set were examined. There were no high inter-correlations among the continuous variables (i.e., $r < 0.5$), but there was a significant association between two item variables: item reference and explicitness of requested information ($X^2 = 7.25$, *df.* = 1, $p < 0.01$). Consequently, only item reference, which has a higher intercoder agreement index (see Table 3), was included in the MLM analyses. The following is a list of the final set of text and item variables included in the MLM analyses:

Item and Item-by-Text Variables

- Question word familiarity
- Item reference to whole or part of text
- Subskill tested
- Number of plausible distractors

Text Variables

- Text length
- Connectives density
- Section
- Nonverbal information

The first step in the MLM analyses was to examine a null model, with no predictors, to identify the proportion of variance in item difficulty that is associated with each level (see column 2 of Table 11). The intercept, which represents average item difficulty across all items and texts, was 0.00. Between-text variance was 0.02 ($X^2 = 69.32$, *df.* = 53, $p = 0.06$), while within text (between-item) variance was 0.54. The Interclass Correlation (ICC) was 0.04, indicating that 4% of the variance in item difficulty is accounted for by differences between texts; the remaining 96% of variance is accounted for by within-text (between-item) variance. Reliability of the intercept was 0.14, indicating that 14% of the variation in the intercept (i.e., average item difficulty) is potentially explicable by text-level predictors.

The next step consisted of assessing the relative importance of each level-1 predictor (i.e., item and item-by-text factor) in accounting for variance in item difficulty estimates. In order to make the interpretation of the intercept easier, subskill, item reference, and number of plausible distractor were uncentered, while question word familiarity was grand mean centered (i.e., the variable mean across all items and texts). In this way, the intercept could be interpreted as the predicted difficulty of an item that refers to the whole text (coded 0), assesses local or inferential understanding (coded 0), has no plausible distractors (i.e., 0 plausible distractors), and has an average value of question word familiarity ($M = 569.89$, see Table 5). *HLM* provides a *t*-test that tests whether, on average, the relationship between a given predictor and the outcome variable is significantly different from zero (Hox, 2002). The results indicated that the coefficients for question word familiarity (-0.01), item reference (0.36), and number of plausible distractors (0.17) were significant ($p < 0.05$), but the coefficient for subskill tested (-0.21) was not ($p > 0.05$). This indicates that the first three variables have a statistically significant association with estimates of item difficulty. Overall, as question word familiarity increases, items become less difficult. Items that refer to part of the text were more difficult than items referring to the whole text. As the number of plausible distractors increases, item difficulty also increases. Model fit statistics indicated that the addition of the four level-1 predictors significantly improved model fit (*X2* = 28.95, *df.* = 4, $p < 0.001$). The addition of the four predictors explained 15% of the variance at level 1 (i.e., within-text variance) in item difficulty estimates.[8]

Next, in order to examine whether the relationships between each of the four level-1 predictors and item difficulty varied significantly across level-2 units (i.e., texts), four models were specified and tested. In each of these models the relationship between the predictor and item difficulty was allowed to vary across texts. Two statistics were examined to assess whether the association between item difficulty and a given level-1 predictor varied significantly across texts: model fit indices (i.e., deviance statistics) and Chi-square tests which test whether a coefficient has a significant random variance across level-2 units (Barkaoui, 2013a; Hox, 2002; Luke, 2004). Only one variable, item reference, showed significant variation in its relationship with item difficulty across texts ($X^2 = 47.36$, *df.* = 31, $p < 0.05$). This suggests that the relationship between whether an item refers to the whole text or part of the text and item difficulty varied significantly across texts.

The next model examined the relationships between level-2 (text) variables and item difficulty estimates. Each of the four text variables was added to the model to find out if it is significantly associated with item difficulty estimates. Only two text variables were found to have significant coefficients at $p < 0.05$: text length (0.002) and section A compared to sections B and C (-0.30). The addition of these two text variables also improved model fit significantly. These results indicate that section A items were significantly easier (by about

[8] The level-1 variance for this model was 0.46, while that for the null model was 0.54.

Table 11: MLM Results for Null and Final Models for Item Difficulty

	Null Model	Final Model
Fixed Effects		
Level 1: Coefficient (SE)		
Intercept (γ_{00})	-0.00 (0.05)	-0.35** (0.10)
Question word familiarity (γ_{10})		-0.01* (0.003)
Question reference (γ_{20})		0.17 (0.13)
Number of plausible distractors (γ_{30})		0.17** (0.05)
Level 2: Text Coefficient (SE)		
Intercept (γ_{01})		
Text length (γ_{21})		0.005** (0.001)
Random Effects		
Between-text Variance (μ_0)	0.023	0.03
X^2 (df, p-value)	69.32 (53, 0.06)	28.38 (31, p > 0.05)
Question reference Variance (μ_2)		0.21
X^2 (df, p-value)		44.57 (30, 0.04)
Within-text Variance (r)	0.54	0.42
ICC	0.04	
Reliability		
Intercept (γ_{00})	0.14	0.18
Question reference (γ_{20})		0.32
Model Fit		
Deviance (#parameters)	487.57 (3)	445.63 (9)
Model Comparison: X^2 (df.)		41.94** (6)

* p < 0.05; ** p < 0.01

a third of a logit) than items in sections B and C. This is in line with the expectations of the test designers. Additionally, on average, item difficulty increased by 0.002 logits with each additional word (or about one fifth of a logit per 100 words), controlling for the effects of the level-1 predictors in the model. The final model aimed to evaluate the effect of the two text variables, text length and section A, on the association between item reference and item difficulty in order to explain the variance in this association across texts. Only text length was found to have a significant effect on the association between item reference and item difficulty (0.005). The addition of text length to the model improved model fit significantly ($X^2 = 7.43$, *df.* = 1, $p < 0.05$). However, the inclusion of text length also reduced the effect of item reference on item difficulty, making it nonsignificant. This suggests that the association between item reference and item difficulty depends significantly on text length. When text length was taken into account, the effects of item reference on item difficulty disappeared. To further examine this interaction effect, the correlations between text length and item difficulty were estimated and compared for items that refer to the whole text versus items that refer to part of the text. Pearson *r* was 0.14 for items that refer to the whole text and 0.40 for items that refer to part of the text. This suggests that the relationship between item difficulty and text length was stronger when the item referred to a specific part of the text. Generally, items that refer to a specific part of longer texts were more difficult than items that refer to a specific part of shorter texts. This is not surprising because when responding to an item that refers to a specific part of a long text, test takers have to process and sift through more details than when responding to an item that refers to a specific part of a short text. The inclusion of text length as a moderator variable also reduced the main effects of text length and section A on item difficulty estimates rendering them nonsignificant.

Based on the results presented above a final MLM model was developed that included three level-1 (item) variables: question word familiarity, item reference, and number of plausible distractors, and one level-2 (text) predictor: text length as moderator of the relationship between item reference and item difficulty estimates. The results of this final model are reported in column 3 of Table 11 above. As Table 11 shows, question word familiarity and number of plausible distractors have significant associations with item difficulty. Overall, as the word familiarity of the item increases, item difficulty decreases. As the number of plausible distractors increases, item difficulty increases by 0.17 logits for each additional plausible distractor. The relationship between question reference and item difficulty is not significant, but this relationship varied significantly across texts depending on text length as explained above. Model fit statistics indicated that the final model fits the data significantly much better than the null model. However, the variables included in the final model explained only 22% of the variance in item difficulty.

Relationships between Text and Item Characteristics and Item Bias Indices

This section reports and discusses findings concerning the relationships between the linguistic and discourse characteristics of MET texts, on one hand, and item bias indices, on the other, to address research question 5. As noted above, items involved in biased interactions were coded 1 (biased); other items were coded 0 (no bias).[9] Table 12 reports descriptive statistics for the continuous text and item variables across item bias status (biased vs. no bias) in relation to L1 and age group, while Table 13 displays the frequencies for categories of each categorical text and item variable across item bias status (biased vs. no bias) in relation to L1 and age group. The means in Table 12 were compared across biased and nonbiased items for L1 and age group using ANOVA, with item bias status as the independent variable and each continuous text and item variable (in Table 12) as the dependent variable. In order to examine the associations between item bias status (biased vs. no bias) and the categorical text and item variables in Table 13, Chi-square (X^2) tests were conducted for item bias status and each categorical variable in Table 13. For L1 group, two text variables (lexical density and AWL) and two item variables (item length and AWL) were found to be significantly (p < 0.05) associated with item bias status, although in all cases effect size was small:

- *Text lexical density* (F[1, 214] = 7.09, p < 0.05, η^2 = 0.03): Texts associated with items exhibiting significantly biased interactions with L1 have lower lexical density (M = 0.57) than texts associated with items not involved in significantly biased interactions with L1 (M = 0.59).
- *Text average word length* (F[1, 214] = 4.24, p < 0.05, η^2 = 0.02): Texts associated with items exhibiting significantly biased interactions with L1 have lower AWL (M = 5.07) than texts associated with items not involved in significantly biased interactions with L1 (M = 4.95).
- *Question length* (F[1, 214] = 8.82, p < 0.05, η^2 = 0.04): Items exhibiting significantly biased interactions with L1 were shorter (M = 30.31 words) than items not involved in significantly biased interactions with L1 (M = 34.15 words).
- *Question average word length* (F[1, 214] = 3.81, p = 0.05, η^2 = 0.02): Items exhibiting significantly biased interactions with L1 had higher AWL (M = 5.01) than items not involved in significantly biased interactions with L1 (M = 4.86).

For age group, two text variables (word polysemy and referential cohesion) and three item variables (item length, item reference, and percentage of words in text relevant to question) were found to be significantly (p < 0.05) associated with item bias status, although in all cases effect size was small:

- *Text word polysemy* (F[1, 214] = 3.86, p = 0.05, η^2 = 0.02): Texts associated with items exhibiting significantly biased interactions with age group have lower word polysemy (M = 3.76) than texts associated with items not involved in significantly biased interactions (M = 3.89).

[9] As Table 8 shows, the same item can be positively biased for one group and negatively biased for another. This distinction is ignored in the following analyses. Each item is coded as being biased (1) or not biased (0), without considering the direction of bias. This approach allows identifying the text and item characteristics which are associated with item bias, but does not allow identifying which variables explain positive vs. negative item bias.

- *Text referential cohesion* (text argument overlap) (F[1, 214] = 3.70, p = 0.05, $\acute{\eta}^2$ = 0.02): Texts associated with items exhibiting significantly biased interactions with age group have lower argument overlap scores (M = 0.36) than texts associated with items not involved in significantly biased interactions with age group (M = 0.42).
- *Question length* (F[1, 214] = 7.63, p < 0.05, $\acute{\eta}^2$ = 0.03): Items exhibiting significantly biased interactions with age group were shorter (M = 30.05 words) than items not involved in significantly biased interactions with age group (M = 33.80 words).
- *Percentage of words in text relevant to question* (F[1, 214]= 5.07, p < 0.05, $\acute{\eta}^2$ = 0.03): Items exhibiting significantly biased interactions with age group had a lower percentage of words in text relevant to question (M = 34.06%) than items not involved in significantly biased interactions with age group (M = 47.02%).
- *Item reference* (X^2 = 8.04, *df.* = 1 p < 0.05): Examination of the cell frequencies showed that half of the items that refer to part of the text (n = 26 out of 52 items, 50%) were involved in significantly biased interactions with age group, while less than third of the items that refer to the whole text (n = 47 out of 164 items, 29%) exhibited significantly biased interactions with age group.

In order to further examine the relationships between MET text and item characteristics, on one hand, and item bias status, on the other, logistic MLM was used, with item bias status (a binary variable) as the outcome. As explained above, logistic MLM is used when the outcome variable is binary (0 or 1) and aims to estimate the contribution of the predictors (continuous and/or categorical) to the probability or likelihood that the outcome is 1 (i.e., item showing significantly biased interaction with a test-taker variable) (Hox, 2002; Luke, 2004). Separate logistic MLM analyses were conducted for L1 and age group. Because the number of item and text variables in the study was large, it was necessary to select only a subset of the item and text variables for inclusion in MLM analyses. Consequently, only variables that were significantly (p < 0.05) associated with item bias status (based on ANOVA and X^2 results above) were selected for inclusion in the logistic MLM analyses. The following is a list of the set of text and item variables included in the logistic MLM analyses for L1 and age group:

	L1 Bias	Age Group Bias
Text	• Lexical density • Average word length	• Text word polysemy • Text referential cohesion
Item	• Question length • Question average word length	• Question length • Percentages of words in text relevant to question • Item reference

The first set of logistic MLM models examined items that were involved in significantly biased interactions with test-taker L1 group. These items (n = 107) were coded 1. The first MLM model assessed the relationship between question length and AWL (both grand-mean centered), on one hand, and item bias for L1, on the other. Question length was found to have a significant coefficient (-0.34, p < 0.05), but not question AWL (-0.04, p > 0.05). Both variables also did not have significant variance coefficients (p > 0.05). The next set of models examined the relationship between level-2 (text) variables and item bias for L1. Each of the two text variables (text lexical density and text AWL) was added to the model to find out if it is significantly associated with item bias for L1. Only one text variable, text lexical density, was found to have a significant coefficient (-8.55, p < 0.05). Based on these results, a final model was specified that included one level-1 (item) predictor, question length, and one level-2 (text) predictor, text lexical density. Both variables were grand-mean centered. A summary of the results of this final model is presented in Table 14.

For bias in relation to L1 group, the intercept in Table 14 (-0.38) is the expected log-odds of the outcome variable being 1 (i.e., an item being involved in a significantly biased interaction with L1) for an item with average word length (i.e., 32.53 words, see Table 5) associated with a text with average lexical density (i.e., 0.58, see Table 4). The coefficients for the two predictors in Table 14 represent the change in the logit of the outcome variable associated with a one unit change in each predictor variable. Negative values indicate a decrease in the likelihood of the outcome being 1

Table 12: Mean Comparisons for Continuous Text and Item Variables across Item Bias Status for L1 and Age Group

	L1 Group				**Age Group**			
Bias Status	**Biased** (*n* = 91)		**Not Biased** (*n* = 125)		**Biased** (*n* = 73)		**Not Biased** (*n* = 143)	
Text variable	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
Length	214.06	78.00	214.65	89.61	209.68	88.63	216.67	80.02
Syntactic complexity								
Sentence length	14.17	4.35	15.21	4.68	14.11	4.54	14.86	4.49
Syntactic similarity	0.09	0.02	0.09	0.02	0.09	0.02	0.09	0.02
Lexical Features								
Density*	0.57	0.05	0.59	0.06	0.57	0.06	0.58	0.05
Variation	101.51	25.12	105.89	27.49	106.96	23.45	101.52	27.35
Sophistication								
lambda	1.53	0.49	1.45	0.44	1.51	0.50	1.49	0.46
AWL*	5.07	0.41	4.95	0.40	5.04	0.41	5.00	0.41
Word information								
Frequency	2.09	0.15	2.12	0.16	2.11	0.16	2.10	0.15
Familiarity	567.30	7.81	568.58	7.48	568.27	7.60	567.62	7.74
Polysemy**	3.84	0.45	3.85	0.42	3.76	0.43	3.89	0.44
Coherence/Cohesion								
Referential**	0.40	0.24	0.41	0.21	0.36	0.22	0.42	0.23
Conceptual	0.22	0.12	0.21	0.14	0.21	0.12	0.22	0.13
Connectives density	82.39	19.70	82.99	21.39	81.12	20.53	83.42	20.33
Pronoun incidence	39.77	27.88	45.08	30.06	40.36	23.48	42.84	31.31
Concreteness	0.27	0.89	0.28	0.92	0.33	0.86	0.24	0.92
Readability								
Flesch Reading Ease	52.17	11.68	54.61	11.39	52.71	11.52	53.45	11.67
Item variables								
Item Length***	30.31	9.68	34.15	9.18	30.05	10.17	33.80	9.01
Item vocabulary								
Item word familiarity	569.76	16.39	570.00	14.86	570.95	15.50	569.36	15.50
Item AWL*	5.01	0.62	4.86	0.51	4.95	0.64	4.91	0.52
Lexical overlap of correct answer & distractors	0.12	0.18	0.11	0.15	0.10	0.16	0.12	0.16
Level of abstractness	2.15	1.17	2.10	1.19	2.01	1.17	2.17	1.18
Percentage of relevant text**	41.73	40.25	43.31	40.65	34.06	37.21	47.02	41.37
Number of plausible distractors	1.55	0.93	1.46	0.85	1.59	0.86	1.45	0.89

* indicates statistically significant differences across item bias status for L1 group

** indicates statistically significant differences across item bias status for age group

*** indicates statistically significant differences across item bias status for both L1 and age group

Table 13 Frequency Comparisons for Categorical Text and Item Variables across Item Bias Status for L1 and Age Group

	L1		Age Group	
Bias Status	**Biased** (*n* = 91)	**Not Biased** (*n* = 125)	**Biased** (*n* = 73)	**Not Biased** (*n* = 143)
Text variables				
Domain				
Public	58	80	50	88
Occupational	30	41	20	51
Educational	3	4	3	4
Section				
A	23	19	18	24
B	21	46	19	48
C	47	60	36	71
Nonverbal information				
Yes	66	82	53	95
No	25	43	20	48
Topic				
Health/Psychology	16	14	12	18
Environment	16	19	9	26
Economic	28	37	24	41
Science & Technology	12	19	7	24
Everyday life	19	36	21	34
Item variables				
Correct answer position				
First	25	27	13	39
Second	28	37	24	41
Third	14	28	14	28
Fourth	24	33	22	35
Number of texts to answer question				
One text	82	110	67	125
Multiple texts	9	15	6	18
Item reference to:**				
Whole text	68	96	47	117
Part of text	23	29	26	26
Subskill tested				
Global	18	26	10	34
Local	38	54	4	75
Inferential	20	28	14	34
Explicitness of requested information:				
Explicit	42	72	34	80
Implicit	49	53	39	63
Location of requested information				
Early	20	15	16	19
Middle	23	30	21	32
Late	18	36	19	35
Entire text	21	29	11	39
Multiple texts	9	15	6	18

** indicates statistically significant differences across item bias status for age group

Table 14: MLM Results for Final Model for L1 and Age Group Bias

L1 Group					
Fixed Effects	*Coefficient*	*SE*	*t-ratio*	*d.f.*	*p*
Intercept	-0.38	0.16	-2.46	52	0.02
Level 1: Question length	-0.05	0.01	-3.43	213	0.001
Level 2: Text lexical density	-8.21	2.85	-2.89	52	0.006
Random Effects	*Var. Comp.*	X^2	*d.f.*	*p*	
Intercept	0.13	56.37	52	0.31	
Age Group					
Fixed Effects	*Coefficient*	*SE*	*t-ratio*	*d.f.*	*p*
Intercept	-0.88	0.17	-5.26	52	0.000
Level 1: Question length	-0.03	.01	-2.53	212	0.01
Level 1: Item reference	0.84	.34	2.46	212	0.01
Level 2: Text word polysemy	-0.88	.43	-1.93	52	0.06
Random Effects	*Var. Comp.*	X^2	*d.f.*	*p*	
Intercept	0.004	60.24	52	0.20	

(i.e., an item being biased for L1). Positive coefficients indicate an increase in the likelihood of the outcome variable being 1. Thus, for each increase of one word in question length, there is a decrease of -0.05 in the log-odds to ([-.38] + [-.05] =) -0.43. For each unit increase in text lexical density, there is a decrease of -8.21 in the log-odds to ([-0.38] + [-8.21] =) -8.59. Overall, it seems that shorter questions (i.e., fewer words) and texts with lower lexical density (i.e., lower proportion of content to function words) tended to be more likely to be associated with bias for L1 than do longer questions and more lexically dense (i.e., contain more content words) texts.

The second set of logistic MLM models examined items involved in significantly biased interactions with test-taker age group. These items (n = 96) were coded 1. A first model assessed the relationship between three item variables (question length, item reference, and percentage words in text needed to answer a question), on one hand, and item bias for age group, on the other. Only two item variables were found to have significant coefficients at $p < 0.05$: question length (-0.03) and item reference (0.75). Neither variable has a significant variance coefficient ($p > 0.05$), however. The next model estimated the relationships between two text measures (text word polysemy and text referential cohesion) and item bias for age group. Only text word polysemy was found to have a significant coefficient (-0.73, $p = 0.05$). Based on these results, a final model was specified that included two level-1 (item) predictors, question length and item reference, and one level-2 (text) predictor, text word polysemy. A summary of the results of the final model is presented in Table 14.

For bias in relation to age group, the intercept in Table 14 (-0.88) is the expected log-odds of the outcome variable being 1 (i.e., an item being involved in a significantly biased interaction with age group) for an item that refers to the whole text (coded 0), has an average word length (i.e., 32.53 words, see Table 5), and is associated with a text with average word polysemy score (i.e., 3.79, see Table 4). For each increase of one word in question length, there is a decrease of -0.03 in the log-odds to ([-0.88] + [-0.03] =) -0.91. For each unit increase in text word polysemy, there is a decrease of -0.88 in the log-odds to ([-0.88] + [-0.88] =) -1.76. Items that refer to a part of the text (coded 1) are associated with a higher likelihood for item bias for age group, with a coefficient of 0.84 which increases the log-odds to ([-0.88] + 0.84 =) -0.04. Apparently, shorter questions, questions that refer to a specific part of the text, and texts with lower word polysemy scores tended to increase the likelihood for an item to be significantly biased in relation to test-taker age compared to longer questions, questions that refer to the whole text, and texts with higher word polysemy scores.

Summary and Discussion

By design, the 54 MET reading texts included in this study varied in terms of length, domain, topic, and whether they included nonverbal information or not. They also varied significantly in terms of sentence length, which is closely associated with text length, but not in terms of other measures of syntactic complexity and lexical characteristics across sections. Additionally, the MET texts included more words that are less familiar, less frequently used, and have fewer senses (i.e., lower polysemy) than the authentic and simplified ESL texts for beginner and intermediate ESL learners examined in other studies. The inclusion of more infrequent words and fewer familiar words makes the texts more difficult to process, while the much lower polysemy values suggest that the MET texts examined in this study included relatively fewer ambiguous lexical items than those in the studies mentioned above.

The MET texts also varied significantly in terms of coherence and cohesion across sections. Overall, it seems that section C texts contain more markers of textual coherence and cohesion than section B texts, which in turn had higher scores on some indices of textual cohesion than section A texts. Once again, this variability might be a function of the variability in text length and type across sections as well. For example, section A texts tend to consist of advertisements/flyers and, because of the nature of that genre, have shorter sentences and fewer connectors than section B and C texts. Finally, the MET texts varied significantly in terms of their readability scores (i.e., Flesch Reading Ease) across sections. Because of the differences across sections in terms of length and other features discussed above, section C texts are significantly more challenging than section A texts. This is, obviously, intended: section C texts are purposefully designed to be more challenging. Still, the MET texts seem to be much easier than IELTS Academic reading texts. This is not surprising given that IELTS targets higher proficiency levels (levels B1 to C2 on CEFR) compared to MET (which targets levels A2 to C1).

The 216 reading items included in the study varied along several dimensions, including the number of texts relevant to the question, subskill tested, whether items refer to the whole text or part of the text, whether the information that the items required was explicitly or implicitly mentioned in the text, the ordinal position of the correct answer, the location of the required information in the text, the number of plausible distractors, item length and vocabulary, and the level of the abstractness of the information required by the question. Some of these item variables varied significantly across sections as well. In particular, items associated with section A texts tended to be significantly shorter (i.e., included fewer words) than those associated with section B and with multiple-text items. Section C items were also significantly shorter than those associated with multiple-text items. Reassuringly, question AWL and word familiarity were very similar to those of the reading texts on which the items are based. However, items in section B tend to have higher word familiarity but lower AWL scores than items associated with section C texts. Unsurprisingly, items about multiple texts tended to request information that is more abstract and to require reading and understanding a significantly larger portion of the text than did items based on single texts. Additionally, because section A texts are much shorter, items associated with them required reading a larger portion of the text than did items associated with the much longer section C texts. The degree of lexical overlap between the correct answer and the distractors and the number of plausible distractors did not vary significantly across sections.

Items assessing inferencing were significantly longer than global and local items. Items assessing global understanding had significantly greater lexical overlap between the correct answer and the distractors than other items. Not surprisingly, items assessing inferencing and global understanding had a significantly higher level of abstractness and required reading a significantly greater portion of the text than did items assessing local understanding.

The items in this study have several positive measurement qualities. First, the 216 items varied significantly in terms of their difficulty; the analyses reliably separated the items into more than 13 levels of difficulty. Second, except for two items, all the items had acceptable fit statistics indicating that the items function as expected by the measurement model. Third, although the test seems to be slightly easy for the group of test takers included in the study, it reliably separated the test takers into several statistically distinct levels in terms of the ability being measured. Finally, there were no significantly biased item-by-test-taker and item-by-gender interactions. However, there were several significantly biased item-by-age and item-by-L1 interactions involving more than a third of the items.

Examination of the relationships between the linguistic and discourse characteristics of MET texts and items, on one hand, and item difficulty estimates, on the other, indicated that four text variables (text length, connectives density, section, and nonverbal information) and five item variables (question word familiarity, item reference, subskill tested, explicitness of information requested, and number of plausible distractors) had significant associations with item difficulty estimates. Specifically, traditional statistical analyses (i.e., Pearson *r* correlation, ANOVA) indicated that items based on longer texts were harder than items based on shorter texts (cf. Gorin & Embreston, 2006; Rupp et al., 2001); items based on section C texts were significantly harder than items based on section A texts; texts with more connectives were associated with harder items; items with more plausible distractors tended to be harder than those with fewer plausible distractors (cf. Rupp et al., 2001); items requesting implicit information were harder than items requesting explicit information (cf. Rupp et al., 2001); items assessing local understanding and inferencing were harder than items assessing global understanding (cf. Rupp et al., 2001); items that refer to part of the text were harder than items referring to the whole text; and as item word familiarity decreases, items become more difficult.

However, MLM analyses, which take into account the nested structure of the data and examine multiple predictors simultaneously, indicated that only question word familiarity and number of plausible distractors have significant associations with item difficulty and that text length had a significant effect on the relationship between item reference and item difficulty. All other text and item variables did not have significant main or interaction effects on item difficulty estimates. These findings indicate that, overall, as the word familiarity of the item increases, item difficulty decreases significantly; as the number of plausible distractors increases, item difficulty increases significantly (cf. Rupp et al., 2001); and that text length moderates the main effect of item reference (to the whole or part of the text) on item difficulty estimates. That is, items that refer to a specific part of longer texts tended to be more difficult than items that refer to a specific part of shorter texts. This is not surprising because when responding to an item that refers to a specific part of a long text, test takers have to process and sift through more details than when responding to an item that refers to a specific part of a short text. Unfortunately, the text and item variables included in the final model explained only 22% of the variance in item difficulty.

Items that were significantly biased in relation to test-taker L1 seem to be associated mainly with two text variables (lexical density and AWL) and two item variables (item length and AWL). Traditional statistical analyses indicated that items exhibiting significantly biased interactions with L1 tended to be shorter, to have higher AWL, and to be associated with texts that have lower lexical density and lower AWL, compared to items not involved in such biased interactions. However, MLM analyses indicated that only question length and text lexical density are significantly associated with item bias for L1. Specifically, it seems that shorter questions (i.e., that have fewer words) and texts with lower lexical density (i.e., lower proportion of content to function words) tended to be more likely to be associated with bias for L1 than do longer questions and more lexically dense texts (i.e., texts that contain more content words).

Items that were significantly biased in relation to test-taker age group seem to be associated mainly with two text variables (word polysemy and referential cohesion) and three item variables (item length, item reference, and percentage of words in text relevant to question). Specifically, traditional statistical analyses indicated that items exhibiting significantly biased interactions with age group tended to be shorter, to refer to part of the text, to have a lower percentage of words in the text relevant to the question, and to be associated with texts that have lower word polysemy and lower referential cohesion (as measured by argument overlap scores), compared to items not involved in such biased interactions. However, MLM analyses indicated that only question length, item reference, and text word polysemy are significantly associated with item bias for age group. Apparently, shorter questions, questions that refer to a specific part of the text, and texts with lower word polysemy scores tended to increase the likelihood for an item to be significantly biased in relation to test-taker age compared to longer questions, questions that refer to the whole text, and texts with higher word polysemy scores.

The findings above indicate that the sample of MET texts and items included in this study exhibited several desirable features that support the validity argument of the MET reading subsection. First, the MET appears to test L2 reading comprehension by including a variety of texts and items that vary in terms of several construct-relevant features. The 54 MET texts in this study varied, as intended, in terms of length, text type, domain, and topic. Also, while the MET texts did not vary significantly in terms of syntactic complexity and lexical features, longer texts contained more markers of textual coherence and cohesion and were more challenging than shorter texts. Second, the MET texts appear to be similar to IELTS Academic texts in terms of some of their lexical characteristics and conceptual cohesion, but tend to have shorter sentences and to be easier than the IELTS Academic reading texts in Green at al. (2010). This is expected given that IELTS targets higher proficiency levels (levels B1 to C2 on CEFR) than does the MET (which targets levels A2 to C1).

Third, the MET includes a number of different types of questions that varied along several, construct-relevant dimensions including number of texts relevant to the question, subskill tested, whether the item refers to the whole text or part of the text, the location of requested information in the text, the number of plausible distractors, and the level of abstractness of information requested by the question. The variety of text and item features that are included in MET and that are construct-relevant ensures that the test captures different aspects of reading comprehension and engages various reading processes and skills such as understanding specific details, understanding vocabulary in context, understanding gist, and synthesizing information from multiple texts. Fourth, reassuringly, the vocabulary level (i.e., AWL and word familiarity) of the items is similar to that of the reading texts on which the items are based.

Fifth, as expected, items based on multiple texts and items assessing global understanding and inferencing tend to request more abstract information and to require reading and understanding a significantly larger portion of the text than did items based on single texts or items assessing local understanding. Furthermore, items based on shorter texts require reading a larger portion of the text than do items associated with longer texts. These findings indicate that, as would be expected by theory, responding to items assessing higher-level processes, such as global understanding and inferencing, requires understanding larger segments of the text and dealing with more abstract ideas compared to responding to items assessing lower-level processes, such as understanding specific details. Sixth, the degree of lexical overlap between correct answer and distractors, a construct-irrelevant factor, and the number of plausible distractors did not vary significantly across sections.

Seventh, the MET items in this study had several positive measurement qualities (i.e., a wide range of item difficulty, acceptable item fit, and no significantly biased interactions with test takers). Including a wide range of item difficulty allows the test to target students with a variety of L2 reading ability levels. Eighth, several construct-relevant text variables (i.e., text length, section or text type, and connectives density) and item variables (i.e., subskill tested, explicitness of information requested by item, item reference, and number of plausible distractors) were significantly associated with item difficulty estimates. Reassuringly, construct-irrelevant factors such as text domain and topic and the position of the correct answer were not significantly associated with item difficulty, which suggests that they do not affect test performance. Finally, most text, item, and item-by-text variables included in the study were not significantly associated with item bias indices. Collectively, these findings suggest that performance on the MET is affected mainly by construct-relevant factors and less by construct-irrelevant factors.

However, the MET texts and items included in this study exhibited also some problematic characteristics that need to be addressed in order to improve the test. First, item length and vocabulary level (i.e., word familiarity and AWL) varied significantly across sections. Item length and degree of lexical overlap between correct answer and distractors also varied significantly across items testing different subskills. Additionally, item word familiarity was significantly associated with item difficulty. Item length, item vocabulary level, and degree of lexical overlap between correct answer and distractors are all irrelevant to the construct being measured by the test (text comprehension) and should be standardized across texts and test forms as much as possible in order to eliminate or reduce their potential effects on item difficulty in future forms of the test. Item word familiarity needs also to be reviewed in order to eliminate or reduce the potential effects of the level of familiarity of the question words on test performance.

Second, several construct-relevant text and item-by-text features, such as text syntactic complexity, text lexical characteristics, text coherence and cohesion, text concreteness, text readability, number of texts needed to answer the question, the location of requested information, percentage of relevant text to answer the question, and the level of abstractness of the question, were not significantly associated with item difficulty. Additionally, the variables included in the study explained only about a fifth of the variance in item difficulty estimates. There are several possible explanations for these findings. First, it is possible that MET item difficulty is not influenced by the text and item-by-text variables listed above. This is not to say that these variables are not important, however. For example, while items based on multiple texts do not seem to differ significantly in terms of difficulty from items based on single texts, it is important to include such items in the test in order to assess whether test takers can comprehend and synthesize information from multiple texts. Second, some variables listed above may not be relevant and/or they may overlap with other variables that were significantly associated with item difficulty. For example, location of information requested by the question may not be relevant because test takers could read the text as many times as they wish, which eliminates the effects of the location of information on short-term memory and, consequently, on item difficulty (Rupp et al., 2001). On the other hand, the percentage of relevant text to answer the question and level of abstractness of question seem to be associated with the subskill tested by the item as noted above. Furthermore, several text characteristics examined in this study (e.g., sentence length, text type or section) were associated with text length. This suggests that the number of words in the texts in this study is an indicator of several other text features (e.g., text type, cohesion) in addition to text length. As noted above, text length has a relatively strong relationship with item difficulty. Third, it is possible that the lack of variability across texts and items in terms of some features measured in this study (e.g., text lexical characteristics) did not allow capturing the association of these variables with item difficulty. Fourth, perhaps the variables included in this study were not sensitive enough to detect the specific text and item-by-text characteristics in the MET that contribute to item difficulty.

Finally, more than a third of the items included in the study exhibited significantly biased item-by-age and item-by-L1 interactions. Some item and text features seem to influence whether an item exhibits bias or not. Specifically, shorter items and questions that refer to a specific part of the text tend to be more difficult for some subgroups of test takers and easier for others depending on test-taker age and L1. Additionally, texts with low lexical density and word polysemy scores tended to increase the likelihood for an item to be significantly biased in relation to test-taker age and L1. It should be noted here that age and L1 are correlated with other factors, such as cognitive development and literacy practices, that could explain the biased interactions between test-taker L1 and age group, on the one hand, and the item and text characteristics listed above, on the other. Future studies could examine the relationships between these correlates and item bias as well as the items that exhibited bias in order to improve their quality and reduce the likelihood of their biased interaction with test-taker characteristics.

Future Research

The study has some limitations. First, the study focused on item and text characteristics without considering reader factors (e.g., reader characteristics) and how they affect test performance. Research clearly shows that reading is an interactive process that involves complex interactions between reader characteristics, goals and contexts, and text and item characteristics (Alderson, 2000; Khalifa & Weir, 2009). Second, the study adopted a task analysis approach in combination with score analyses, and did not examine the actual processes that test takers engage in when interacting with and responding to the MET reading texts and items. Examining test takers' reading processes could also help explain findings concerning item bias in relation to test-taker L1 and age group. Third, the sample of texts and items included in the study was small. The study could be replicated with a larger sample of texts and items and other groups of test takers. Fourth, as noted above, the variables included in the study did not explain all the variance in item difficulty estimates. Other variables need to be considered, including human ratings of text discourse and content features such as text coherence. Fifth, the study did not consider the effects of interactions among test-taker variables, such as L1 and age, or the effects of correlates of age and L1, such as test-taker cognitive and literacy development, on item bias. It is possible, for example, that test takers from the same L1 but different ages (or at the same age but from different L1s) responded differently to the items in this study. Examining such complex interactions requires larger samples of test takers. Finally, this study was correlational. Experimental studies that manipulate specific text and item characteristics and examine their effects on item difficulty could provide more convincing evidence concerning the relationships between specific text and item characteristics and performance on L2 reading tests (cf. Gorin & Embreston, 2006).

Future studies could also use eye-tracking and/or think-aloud protocols to examine the cognitive and metacognitive processes and strategies that test takers engage when interacting with and responding to the MET texts and items (cf. Anderson et al., 1991; Cohen & Upton, 2007; Gao, 2006) in order to establish whether the test activates the types of mental processes that a theory of L2 knowledge and performance views as essential elements of L2 reading performance (Chapelle, 2008; Cohen, 2012; Weir, 2005). These studies could also examine how these processes and strategies vary depending on item difficulty and text and item characteristics. For example, it is necessary to examine whether and how test takers' reading and response processes vary across sections A, B and C and across items based on single texts and items based on multiple texts. Such studies could provide important evidence concerning the explanation inference of the test's validity argument (Chapelle, 2008). Other studies could compare the characteristics of the MET reading texts and items to the characteristics of reading texts and tasks in target language use situations (cf. Green et al., 2010). It is also important to examine the extent to which the MET texts and items engage test takers in the same cognitive processes involved in reading in real-life contexts in order to evaluate the extrapolation inference of the test's validity argument (Chapelle, 2008; Cohen, 2012).

Finally, this study demonstrates how to combine task and score analyses in order to examine important questions concerning the validity argument of L2 reading tests. Findings from this line of research can provide important validity evidence; supply useful information for developing and improving L2 reading texts and items; and significantly enhance our understanding of the effects of various text and item features on performance on L2 reading comprehension tests.

References

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

Alderson, J. C., Figueras, N., Kuijper, H., Nold, G., Takala, S., & Tardieu, C. (2006). Analysing tests of reading and listening in relation to the Common European Framework of Reference: The experience of the Dutch CEFR construct project. *Language Assessment Quarterly*, *3*(1), 3–30.

Anderson, N., Bachman, L., Perkins, K., & Cohen, A. (1991). An exploratory study into the construct validity of a reading comprehension test: Triangulation of data sources. *Language Testing*, *8*(1), 41–66.

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

Barkaoui, K. (2013a). An introduction to multi-faceted Rasch models. In A. Kunnan (Ed.), *The companion to language assessment* (pp. 1301–1322). Wiley-Blackwell.

Barkaoui, K. (2013b). An introduction to multilevel modeling in language assessment research. *Language Assessment Quarterly*, *10*(3), 241–273.

Bond, T. G., & Fox, C. M. (2007). *Applying the Rasch model: Fundamental measurement in the human sciences* (2nd ed.). Mahwah. NJ: Lawrence Erlbaum.

Buck, G., Tatsuoka, K., and Kostin, I. (1997). The subskills of reading: Rule-space analysis of a multiple-choice test of second language reading comprehension. *Language Learning*, *47*(3), 423–466.

CaMLA (2012). *Michigan English Test Information Pack*. Ann Arbor, MI: Cambridge Michigan Language Assessments.

CaMLA (2014). *MET 2013 Report.* Ann Arbor, MI: Cambridge Michigan Language Assessments.

Chapelle, C. A. (2008). The TOEFL validity argument. In C. A. Chapelle, M. K. Enright, & J. M. Jamieson, *Building a validty argument for the Test of English as a Foreign Language* (pp. 319–352). New York: Routledge.

Cohen, A. D. (2012). Test-taking strategies and task design. In G. Fulcher & F. Davidson (Eds.), *The Routledge handbook of language testing* (pp. 262–277). New York: Routledge.

Cohen, A. D., & Upton, T. A. (2007). I want to go back to the text: Response strategies on the reading subtest of the new TOEFL. *Language Testing*, *24*(2), 209–250.

Crossley, S. A., Allen, D. B., & McNamara, D. S. (2011). Text readability and intuitive simplification: A comparison of readability formulas. *Reading in a Foreign Language*, *23*(1), 84–101.

Crossley, S. A., & McNamara, D. S. (2008). Assessing second language reading texts at the intermediate level: An approximate replication of Crossley, Louwerse, McCarthy, and McNamara (2007). *Language Teaching*, *41*(3), 409–429.

Crossley, S. A., Greenfield, J., & McNamara, D. S. (2008). Assessing text readability using psycholinguistic indices. *TESOL Quarterly*, *42*(3), 475–493.

Crossley, S. A., Louwerse, M., McCarthy, P. M., & McNamara, D. S. (2007). A linguistic analysis of simplified and authentic texts. *Modern Language Journal*, *91*(1), 15–30.

Davey, B. (1988). Factors affecting the difficulty of reading comprehension items for successful and unsuccessful readers. *Experimental Education*, *56*(2), 67–76.

Davey, B. & Lasasso, C. (1984). The interaction of reader and task factors in the assessment of reading comprehension. *Experimental Education*, *52*(4), 199–206.

Dávid, G. (2007). Investigating the performance of alternative types of grammar items. *Language Testing*, *24*(1), 65–97.

Embretson, S. E., & Wetzel, C. D. (1987). Component latent trait models for paragraph comprehension. *Applied Psychological Measurement*, *11*(2), 175–193.

Enright, M. K., Grabe, W., Koda, K., Mosenthal, P., Mulcahy-Ernt, P., & Schedel, M. (2000). *TOEFL 2000 reading framework: A working paper*. TOEFL Monograph MS-16. Princeton, NJ: Educational Testing Service.

Fellbaum, C. (1998). *WordNet: An electronic lexical database*. Cambridge, MA: The MIT Press.

Field, A. (2009). *Discovering statistics using SPSS* (3rd ed.). Thousand Oaks, CA: Sage.

Flesch, R. (1948). A new readability yardstick. *Journal of Applied Psychology*, *32*, 221–233.

Foltz, P.W., Kintsch, W., & Landauer, T.K. (1998). The measurement of textual coherence with Latent Semantic Analysis. *Discourse Processes*, *25*(2&3), 285–307.

Freedle, R. & Fellbaum, C. (1987). An exploratory study of the relative difficulty of TOEFL's listening comprehension items. In R. Freedle & R. Duran (Ed.), *Cognitive and linguistic analysis of test performance*. Norwood, NJ: Ablex.

Freedle, R. & Kostin, I. (1993). The prediction of TOEFL reading item difficulty: Implications for construct validity. *Language Testing*, *10*(2), 133–170.

Gao, L. (2006). Toward a cognitive processing model of MELAB reading test item performance. *Spann Fellow Working Papers in Second or Foreign Language Assessment*, *4*, 1–39.

Gilhooly, K. J. & Logie, R. H. (1980). Age-of-acquisition, imagery, concreteness, familiarity, and ambiguity measures for 1,944 words. *Behavior Research Methods & Instruction*, *12*(4), 395–427.

Gorin, J. S. & Embretson, S. E. (2006). Item difficulty modeling at paragraph comprehension items. *Applied Psychological Measurement*, *30*(5), 394–411.

Graesser, A. C., McNamara, D. S., Louwerse, M., & Cai, Z. (2004). Coh-Metrix: Analysis of text on cohesion and language. *Behavior Research Methods, Instruments, & Computers*, *36*(2), 193–202.

Graesser, A. C., McNamara, D. S., & Kulikowich, J. M. (2011). Coh-Metrix: Providing multilevel analyses of text characteristics. *Educational Researcher*, *40*(5), 223–234.

Green, A., Ünaldi, A., & Weir, C. (2010). Empiricism versus connoisseurship: Establishing the appropriacy of texts in tests of academic reading. *Language Testing* *27*(2), 191–211.

Green, K. (1984). Effects of item characteristics on multiple-choice item difficulty. *Educational and Psychological Measurement*, *44*(3), 551–561.

Hare, V., Rabinowitz, M., & Schieble, K. (1989). Text effects on main idea comprehension. *Reading Research Quarterly*, *24*(1), 72–88.

Hox, J. J. (2002). *Multilevel analysis: Techniques and applications*. Mahwah, NJ: Lawrence Erlbaum.

In'nami, Y. & Koizumi, R. (2009). A meta-analysis of test format effects on reading and listening test performance: Focus on multiple-choice and open-ended formats. *Language Testing*, *26*(2), 219–244.

Khalifa, H. & Weir, C. (2009). *Examining reading: Research and practice in assessing second language reading (Studies in Language Testing Vol. 29)*. Cambridge: UCLES/Cambridge University Press.

Kintsch, W. & Yarbrough, J. C. (1982). Role of rhetorical structure in text comprehension. *Journal of Educational Psychology*, *74*(6), 828–834.

Kirsch, I. S. & Guthrie, J. T. (1980). Construct validity of functional reading tests. *Journal of Educational Measurement*, *17*(2), 81–93.

Kobayashi, M. (2002). Method effects on reading comprehension test performance: Text organization and response format. *Language Testing*, *19*(2), 193–220.

Koizumi, R. (2012). Relationships between text length and lexical diversity measures: Can we use short texts of less than 100 tokens? *Vocabulary Learning and Instruction*, *1*(1), 60–69.

Kondo-Brown, K. (2002). A FACETS analysis of rater bias in measuring Japanese second language writing performance. *Language Testing*, *19*(1), 3–31.

Laufer, B., & Nation, P. (1995). Vocabulary size and use: Lexical richness in L2 written production. *Applied Linguistics*, *16*(3), 307–322.

Linacre, J. M. (2011). *A user's guide to FACETS Rasch model computer program*. Available on line at: www.winsteps.com

Lu, X. (2011). A corpus-based evaluation of syntactic complexity measures as indices of college-level ESL writers' language development. *TESOL Quarterly*, *45*(1), 36–62.

Luke, D. (2004). *Multilevel modeling*. Thousand Oaks, CA: Sage.

Malvern, D., & Richards, B. (2002). Investigating accommodation in language proficiency interviews using a new measure of lexical diversity. *Language Testing*, *19*(1), 85–104.

McCarthy, P. M. & Jarvis, S. (2010). MTLD, vocd-D, and HD-D: A validation study of sophisticated approaches to lexical diversity assessment. *Behavioral Research Methods*, *42*(2), 381–392.

McNamara, D. S., Crossley, S. A., & McCarthy, P. M. (2010). Linguistic features of writing quality. *Written Communication*, *27*(1), 57–86.

McNamara, D.S., Cai, Z., & Louwerse, M.M. (2007). Optimizing LSA measures of cohesion. In T.K. Landauer, D.S. McNamara, S. Dennis, & W. Kintsch (Eds.), *Handbook of latent semantic analysis* (pp. 379–400). Mahwah, NJ: Erlbaum.

McNamara, T. (1996). *Measuring second language performance*. London, UK: Longman.

Meara, P & Bell, H (2001) P_Lex A simple and effective way of describing the lexical characteristics of short L2 texts. *Prospect*, 16, 5–24.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed.) (pp. 13–103). NY: Macmillan.

Mosenthal, P. (1996). Understanding the strategies of document literacy and their conditions of use. *Journal of Educational Psychology*, *88*(2), 314–332.

Ozuru, Y., Rowe, M., O'Reilly, T., McNamara, D. S. (2008). Where's the difficulty in standardized reading tests: The passage or the question? *Behavior Research Methods*, *40*(4), 1001–1015.

Raudenbush, S.W., Bryk, A.S., Cheong, Y.F., & Congdon, R. (2004). *HLM6: Hierarchical linear and nonlinear modeling*. Lincolnwood, IL: Scientific Software International.

Read, J. (2005). Applying lexical statistics to the IELTS speaking test. *Research Notes*, *20*, 10–16.

Rupp, A. A., Garcia, P., & Jamieson, J. (2001). Combining multiple regression and CART to understand difficulty in second language reading and listening comprehension test items. *International Journal of Testing*, *1*(3–4), 185–216.

Spelberg, H. C. L., de Boer, P., & van den Bos, K. P. (2000). Item type comparisons of language comprehension tests. *Language Testing*, *17*(3), 311–322.

Toglia, M. P. & Battig, W. R. (1978). *Handbook of semantic word norms*. Hillsdale, NJ: Lawrence Erlbaum.

Uiterwijk, H. & Vallen, T. (2005). Linguistic sources of item bias for second generation immigrants in Dutch tests. *Language Testing*, *22*(2), 211–234.

Weir, C.J. (2005). *Language testing and validation: An evidence-based approach*. New York, NY: Palgrave Macmillan.