GEPT® LTTC 全民英檢

Language unfolds worlds.
Testing sets standards.

LTTC-GEPT Research Reports RG-06

A Comparability Study between the General English Proficiency Test – Advanced and the Internet-Based Test of English as a Foreign Language (iBT TOEFL)

Antony John Kunnan
Nathan Carr

LTTC® 財團法人語言訓練測驗中心
THE LANGUAGE TRAINING & TESTING CENTER

A Comparability Study between the General English Proficiency Test - Advanced and the Internet-Based Test of English as a Foreign Language (iBT TOEFL)

LTTC-GEPT Research Reports
RG-06

Antony John Kunnan
Nathan Carr

This study was funded and supported by the Language Training & Testing Center (LTTC) under the LTTC-GEPT Research Grants Program 2013-2014

LTTC-GEPT Research Reports RG-06
A Comparability Study between the General English Proficiency Test - Advanced and the Internet-Based Test of English as a Foreign Language (iBT TOEFL)

Published by The Language Training and Testing Center
No.170, Sec. 2, Xinhai Rd., Daan Dist., Taipei, 10663 Taiwan (R.O.C)

First published August 2015

Foreword

We have great pleasure in publishing this report: *LTTC-GEPT Research Reports RG-06*. The study described in this report was funded by the 2013-2014 LTTC-GEPT Research Grants. Headed by Professor Antony John Kunnan of California State University, Los Angeles, USA, the study investigated the comparability of two English language proficiency tests - the GEPT Advanced and the Internet-Based Test of English as a Foreign Language (iBT TOEFL). The study provides validity and reliability evidence for the GEPT Advanced and relate to the concept of *portability of testes* in the use of the Common European Framework of Reference (CEFR).

The GEPT, developed more than a decade ago by the LTTC to serve as a fair and reliable testing system for EFL learners, has gained wide recognition in Taiwan and abroad. It has generated positive washback effects on English education in Taiwan. As the GEPT has successfully reached out to the international academic community with remarkable success over the years, numerous studies and research projects on GEPT-related subjects have been conducted and published as technical monographs, conference papers, and refereed articles in books and journals. In view of the growing scholarly attention on the GEPT, and in order to assist external researchers to conduct quality research on topics related to the test, the LTTC has set up the LTTC-GEPT Research Grants Program, which offers funding to outstanding research projects.

The annual call for research proposals is publicized every October, attracting proposals from all over the world. A review board, which comprises scholars and experts in English language teaching and testing from Taiwan and abroad, evaluates the research proposals in terms of the following criteria:

- the relevance to identified areas of research
- the benefit of the research outcomes to the GEPT
- the theoretical framework, aims and objectives, and methodology of the proposed research
- the qualifications and experience of the research team
- the capability of the research outcomes to be presented at international conferences and published in journals
- the timeline and cost effectiveness of the proposed research

Complete and up-to-date information about the GEPT is available at https://www.lttc.ntu.edu.tw/E_LTTC/E_GEPT.htm. Full research reports can be downloaded at https://www.lttc.ntu.edu.tw/lttc-gept-grants.htm.

We believe that with the further contributions from the external research community, the GEPT will continue to refine its quality and achieve wider recognition at home and overseas.

Hsien-hao Liao
Executive Director
LTTC

Author Biodata

Antony John Kunnan (Ph.D., UCLA) has taught courses in language assessment, research methods and for 25 years in California, Taiwan, Hong Kong and Singapore. He is the author of 18 authored and edited books, over 60 book chapters and journal articles. He has completed numerous research projects for Cambridge English Language Assessment, U.K., U.S. Agency for International Development, Washington, D.C., Language Training and Testing Centre, Taiwan, and Educational Testing Service, Princeton. He was the founding editor of *Language Assessment Quarterly* and past president of the International Language Testing Association. For more information, please visit: www.antonykunnan.com.

Nathan Carr (Ph.D., UCLA) has taught courses in language assessment, research methods and statistics for language assessment for 15 years in California State University, Fullerton. He has many journal articles in Language Testing and Language Assessment Quarterly. He previously completed projects for Oxford University Press and English First. His latest book is *Designing and analyzing language tests*. Oxford: Oxford University Press.

摘要

◆ 研究團隊與研究目的

本研究由美國加州州立大學（California State University）Professor Antony John Kunnan 主持，目的是比較 GEPT 高級測驗與托福 iBT 測驗的試題內容與受試者的成績表現，研究結果可為 GEPT 高級測驗提供信度與效度證據。

◆ 研究問題

1. GEPT 高級閱讀與寫作測驗與托福 iBT 試題內容的異同。
2. 受試者於 GEPT 高級和托福 iBT 閱讀與寫作測驗的表現差異。
3. GEPT 高級和托福 iBT 閱讀與寫作測驗的可比性。

◆ 研究方法摘要

1. 本研究分美國與台灣兩地進行，全部受試者都在取得托福 iBT 的成績後半年內參加 GEPT 高級閱讀與寫作測驗。
2. 資料分析主要有兩部分：(1)試題內容與(2)測驗成績。試題內容是採用 Coh-Metrix 與 LexTutor 的方法，對於測驗文章的段落、長度、主題、測驗重點等面向進行分析；測驗成績的比較則是採用相關性分析（correlation analysis）、探索性與驗證性因素分析（exploratory and confirmatory factor analyses）、迴歸分析（regression analysis）等統計方法進行。

◆ 研究結果摘要

試題內容：

1 探索性和驗證性因素分析結果顯示兩測驗是評量相同的能力（閱讀能力與寫作能力），但對於構念（construct）的重點不同。

2 兩測驗的文章內容許多部份類似，但涵蓋的構念（construct coverage）、試題範圍（item scope）、題型設計（task format）不盡相同。在涵蓋的構念方面，兩測驗都有關於文章細節的題型，但 GEPT 高級測驗使用得更廣泛；托福 iBT 測驗較多主旨大意題，並且有專考詞彙或是從情境推測單字意思的題目；GEPT 高級測驗有測試略讀（skimming）與掃描（scanning）的能力，但托福 iBT 沒有。在試題範圍方面，GEPT 高級測驗需從較廣的文章範圍來找出答題的線索。在題型設計方面，托福 iBT 僅含選擇題，而 GEPT 高級測驗題型較多元，除了選擇題外，還有簡答題與配合題。

測驗成績：

3 測驗成績分析結果顯示：(1) GEPT 高級測驗具有良好的信度；(2) GEPT 高級測驗與托福 iBT 的測驗結果呈現中高度正相關（兩者閱讀的相關係數為.467，寫作為.385）；(3) GEPT 高級測驗的難度較托福 iBT 難（受試者於 GEPT 高級閱讀測驗與托福 iBT 閱讀的平均答對率分別為為 57.9%與 82.9%，GEPT 高級寫作測驗與托福 iBT 寫作測驗分別為 51.1%與 80.1%）。

4 GEPT 高級與托福 iBT 成績的迴歸分析結果顯示 GEPT 高級閱讀達到 68 分（相當於托福 iBT 24 分）與寫作達到 3 級分（相當於托福 iBT 24 分）約等同於 CEFR C1 級數。

Abstract

This study investigated the comparability of two English language proficiency tests - the General English Proficiency Test - A (GEPT-A) and the Internet-Based Test of English as a Foreign Language (iBT TOEFL). Data was collected from test takers in both Taiwan and the United States. The instruments used were item-level participant test performance response data from the GEPT-A reading section, scores on GEPT-A Writing Task 1, iBT reading, writing, speaking, and listening scaled scores, and participant responses to a background information survey and five questions involving their perceptions of the GEPT-A.

Two specific analyses were conducted: First, a content analysis was performed on the passages in the GEPT-A form and on the sample iBT reading passages published by Educational Testing Services, Princeton, for test preparation purposes. This analysis included using Coh-Metrix (Coh-Metrix, n.d.) and LexTutor (Cobb, n.d.) to analyze the cohesion, syntax, and vocabulary used in passages. Further, a task analysis of the construct coverage, scope, and task formats used in the reading comprehension questions on the two tests was also conducted. Second, an analysis of the participant responses on the tests from the two tests was also conducted.

The results of the text analysis showed the reading passages on the two tests are comparable in many ways but differ in several key regards. The task analysis revealed that the construct coverage, item scope, and task formats of the two tests are clearly distinct. In particular, GEPT-A does not include items targeting vocabulary knowledge or the ability to infer the meaning of unfamiliar vocabulary from context, while the iBT includes both, particularly the former. The GEPT-A included only one careful reading main idea item (as opposed to skimming and identifying the main idea of a paragraph), while the iBT replaced them with items requiring test takers to read for major points. The GEPT-A did not include any items that appeared to call for inferencing. On the other hand, both tests assess the ability to read for specific details, although the GEPT-A uses them much more extensively. The GEPT-A included extensive coverage of skimming and scanning, whereas the iBT did not assess these abilities at all. The reading sections of the two tests also differed markedly in terms of item scope, with GEPT-A reading generally requiring test takers to extract the necessary information from a larger portion of the test than did the iBT. Regarding task format, the iBT reading section relied entirely on selected response items, primarily multiple choice, while the GEPT-A made heavy use of short-answer and matching questions in addition to multiple choice.

Analysis of participant test responses indicated that the GEPT-A has good reliability, and that reading comprehension items tend to function quite well. Scores on the GEPT-A and iBT are highly inter-correlated with each other: More specifically, scores on the two reading tests had a medium-to-large correlation (r = .467), and scores on the GEPT-A Writing Task 1 and iBT writing had a medium-sized correlation (r = .385). The GEPT-A may be somewhat more difficult than the iBT, given that the mean GEPT-A reading score was 57.9% of the total points possible, vs. 82.9% for the iBT reading section, and 51.1% for GEPT-A Writing Task 1 vs. 80.1% for the iBT writing section; however, these comparisons across differently scaled scores must be interpreted with caution. Regression analyses indicated that GEPT-A reading score of 68 corresponded to C1 on the CEFR, by virtue of equating iBT reading scores. Similarly, GEPT-A writing task 1 score of 3 corresponded to C1 on the CEFR.

Exploratory and confirmatory factor analyses of the test score data indicated that the two tests appeared to both be measuring reading and writing ability, but emphasize different aspects of the reading construct—that is, the different construct definitions for the two tests are reflected in the results of the factor analyses. Results on the two tests are therefore not entirely comparable.

Table of Contents

Table of Figures

Table of Tables

1. Introduction

This study was an investigation into the comparability of two language tests, the General English Proficiency Test – Advanced (GEPT-A hereafter) and the Internet-Based TOEFL (iBT hereafter). This is a meaningful study, as the two tests have a similar purpose and a similar test taker customer-base. Both tests are used for admission and selection purpose to undergraduate and graduate programs in U.S. and Canadian English-medium universities. Obviously, this superficial similarity does not necessarily mean that the tests are comparable. This study investigated in depth the degree to which the two tests are in fact comparable, and the degree to which they measured similar language abilities (with a focus on reading and writing). The findings from the study relate to the concept of *portability of tests* listed as one of the key components in the use of the Common European Framework of Reference (CEFR). Portability in the CEFR refers to the use of a particular test in lieu of another when both tests are available (Kunnan, 2012).

2. Review of the Relevant Literature

2.1. The Cambridge-TOEFL Comparability Study

The most widely-known comparability study of language tests was conducted in 1987-90 and published in 1995 by Lyle Bachman and colleagues. It investigated the comparability of the *First Certificate of English (FCE)* administered by the University of Cambridge Local Examinations Syndicate (now Cambridge ESOL) and the paper-and-pencil version of the *Test of English as a Foreign Language (TOEFL)* administered by Educational Testing Service, Princeton.

The research steps used were two-fold: a qualitative content analysis of the items, tasks and prompts and a quantitative analysis of the test performance of the subjects on the tests. The most important aspect of the study related to the care taken in choosing the test instruments, test 3 samples (in terms of test taker characteristics and test takers' score norms) and administrative and scoring procedures in such a way that they truly represented the two testing practices.

The instruments used were authentic *FCE* tests and institutional (retired) *TOEFL* and *SPEAK*, the retired form of *The Test of Spoken English*. But as the institutional version of the *Test of Written English* was not available, a similar test was developed by the study researchers. Sampling procedures included selecting subjects that represented the characteristics of the test takers of the two tests. In addition, an examination of the descriptive statistics demonstrated that the means and standard deviations of the study subjects and the test norms of world-wide test taker groups for the two tests were only two points apart and were not practically different. Further, the administration and scoring procedures mirrored the procedures used by the test administrators and raters of the two tests. With these strict measures in place, it was possible to make conclusions regarding comparability based on the test content analyses, the test performance analyses, and the correlational analyses.

The qualitative content analysis of the two tests was conducted by expert judges who used the Communicative Language Ability instrument developed for the study. It was concluded that in general there were more similarities between the two tests than there were differences. The quantitative statistical analysis of the two tests was conducted by analyzing the test performances of the study subjects. The procedures included descriptive statistics, reliability analyses, correlational analyses and factor structure analyses for each of the individual tests and then across the two tests. The study concluded that as the same higher-order factor structure was supported individually for the two tests and across the two tests, the two tests generally measured similar language abilities.

2.2. Other Comparability Studies

The content analyses conducted in Bachman et al. (1995) resulted in the development and use of the Communicative Language Ability (CLA) and Test Method Facet (TMF) frameworks. This part of the study was published in Bachman, Davidson, and Milanovic (1996). It outlined the need for the framework and a systematic procedure to analyze test content – the linguistic characteristics, test format characteristics and the communicative language abilities that were tested in the test items, tasks and prompts. Another study, Kunnan (1995), from the same dataset focused on test taker characteristics and test performance on the two tests, the *FCE* and the *TOEFL*. This study conducted exploratory and confirmatory factor analyses and structural equation modeling on the test performance and it once again showed that the two tests measured similar language abilities.

Another type of study that compared two versions of the same test was the Choi, Kim, & Boo study (2003). In this study a paper-based language test and a computer-based language test - the *Test of English Proficiency* developed at Seoul National University. The findings based on content analysis using corpus linguistic techniques supported the notion that both versions of the test are comparable. Along similar lines, a study investigated the comparability of conventional and computerized tests of reading in a second language (Sawaki, 2001). This study was a comprehensive review of literature regarding cognitive ability, ergonomics, education, psychology, and L1 reading. The study did not draw clear conclusions and generalizations about computerized language assessments due to the range of characteristics such as administrative conditions, computer requirements, test completion time, and test takers' effect. Similarly, yet another study investigated the comparability of direct and semi-direct speaking test versions 4 (O'Loughlin, 1997). In general, the author concluded that the live and tape-based versions of the oral interaction sub-tests cannot be substituted for each other.

A different approach to investigating equivalence focused on the psychometric equivalence between tests was conducted by Geranpayeh (1994). The study investigated the comparability of scores of subjects who took the TOEFL and the International English Language Testing Service (IELTS). The study found high to moderate correlations between the TOEFL and IELTS scores. The focus on the study was score equivalence: what did 600 on the TOEFL mean on the IELTS?

Along the same lines, Bridgeman and Cooper (1998) conducted a study to investigate the comparability of scores from hand-written and word-processed essays. Results indicated that the scores were higher on the hand-written essays than on word-processed essays. Other investigations included: Weir and Wu (2006) investigated the task comparability in semi-directed speaking tests of three forms of the GEPT-Intermediate. Similarly, Stansfield and Kenyon (1992) investigated the comparability of the oral proficiency interview and the simulated oral proficiency interview. Hawkey (2009) compared historical issues and themes in developing the two tests from Cambridge ESOL: the *First Certificate of English* and the *Certificate in Advanced English*.

Other related procedures have included developing concordance tables of references comparing two or more tests. These procedures only use test scores in developing the tables and are, therefore, not rigorous and do not provide sufficient evidence for comparability of tests.

2.3. Summary

The two Bachman et al. studies (1995, 1996) obviously have direct bearing on the proposed study as these studies investigated the comparability of two different tests. The other studies were investigations of different versions of the same test. In addition, they generally only focused on score equivalence, and this is insufficient data to reach conclusions regarding comparability of tests. Therefore, this study used principles from the Bachman studies such as carefully choosing test instruments, test samples (in terms of test taker characteristics) and administrative and scoring procedures in such a way that they truly represent testing practices of the two tests that are being compared, and conducted both a test content and test performance analysis.

3. Research Questions

Based on the previous work done in this area, and the goals of this project, the following research questions were proposed for this study:

1. What is the content of the reading and writing tests of the GEPT-A and the iBT – the test passages, items, tasks?

2. What is the test performance of the study participants on the reading and writing tests of the GEPT-A at the total test score and at the item level, and on the iBT at the total test score level?

3. What is the comparability of the reading and writing tests of the GEPT-A and iBT?

4. Method

4.1. Participants

The participants in this study consisted of 186 test takers from two groups, one group recruited in Taiwan and the other group recruited in the United States. The Taiwan sample included 118 participants recruited by LTTC staff, but only 92 could be used in the study, because overall scores and demographic data were not available for the other 26 (i.e., there were no test or survey responses for those cases).

The U.S. sample consisted of 92 international students from three American universities: California State University, Fullerton (n = 28); Indiana University, Bloomington (n = 50); and the University of Texas at San Antonio (n = 14). All participants had previously taken the iBT. The U.S. participants were recruited by faculty on the three campuses, using a combination of e-mail announcements, flyers, and word of mouth recruitment. As an incentive, U.S.

participants received a $50 gift card to a coffee shop chain or to their university bookstore after returning their completed tests and surveys and copies of their iBT score reports.

4.2. Instruments

GEPT – Advanced Reading and *Writing.* The primary instrument for this study was one form of the GEPT-A Reading Comprehension Test (Form AR-1101) and Advanced Level Writing Test. The reading test consisted of two sections. The first section, *Careful Reading*, included 4 passages and 20 items, including multiple-choice, short answer, and matching items. The expected responses for the short answer questions varied from one word to one or two sentences. The second section, *Skimming and Scanning*, included 3 passages (with the last passage consisting of 3 shorter sub-passages) and 20 items. The items included both matching and fixed-frame multiple choice (i.e., the options were the same for all multiple-choice questions).

The writing test consisted of two tasks: Task 1 required test takers to read two essays and write an essay in response. Task 2 required students to write a letter to the editor reacting to the information contained in two graphs. Only scores from Task 1 were used in the study.

Test takers survey. The study participants took a brief survey after taking the GEPT-A. The survey asked several background information questions, inquired about test takers' iBT test dates and scores on the four test sections, and included six Likert-scale items about the GEPT-A. The items asked test takers about the difficulty of the GEPT-A reading and writing sections, and about the relevance of the test content and test tasks to their academic studies.

Internet-Based TOEFL (iBT). Three practice iBT reading and writing tests were used for the content comparison of the GEPT-A and iBT. These iBT test forms were taken from *The Official Guide to the TOEFL Test* (Educational Testing Service, 2012). These forms should be highly comparable to the current operational iBT. Each reading test featured 3 passages and 38-42 items, including multiple-choice, a limited number of multiple response multiple choice (MRMC) questions (selected response items in which test takers must select two or more correct options per item), and one categorization item. The writing tests each contained one integrated writing task and one independent writing task. The integrated writing tasks required test takers to read a short text and listen to a short recording before writing an essay. The independent writing tasks required test takers to write an essay in response to a brief prompt. Table 1 lists the passages used in the study.

The study participants were also required to submit copies of their iBT score reports in order to avoid problems with inaccurately remembering their scores.

Table 1. Reading and Listening Passage Identification Key for GEPT-A and iBT

Passage Number	Topic	Passage Number	Topic
GEPT R1	Caravaggio	iBT R1	19th Century Politics in the United States
GEPT R2	Value-Added Assessment	iBT R2	The Expression of Emotions
GEPT R3	Hydrates	iBT R3	Geology and Landscape
GEPT R4	Brownfields	iBT R4	Feeding Habits of East African Herbivores
GEPT R4a	Brownfields summary paragraph	iBT R5	Loie Fuller
GEPT R5	Hudson's Bay Company	iBT R6	Green Icebergs
GEPT R6	Victor the Wild Child	iBT R7	Architecture
GEPT R7 [b]	Three Historical Attractions [a]	iBT R8	Long-Term Stability of Ecosystems
GEPT R7a	Colonial Williamsburg	iBT R9	Depletion of the Ogallala Aquifer
GEPT R7b	Historical Village of Hokkaido	iBT WR1	Altruism
GEPT R7c	Rothenburg ob der Tauber	iBT WL1	Altruism
GEPT WR1	Online Reviews: A Boon for Travelers and Businesses	iBT WR2	Professors on Television
GEPT WR2	The Downside of Online Travel Reviews	iBT WL2	Professors on Television
		iBT WR3	Portrait of an Elderly Woman in a White Bonnet
		iBT WL3	Portrait of an Elderly Woman in a White Bonnet

[a] Colonial Williamsburg, Historical Village of Hokkaido, and Rothenburg ob der Tauber.

[b] Headings for the three sub-passages were included in the Coh-Metrix and vocabulary analyses of the combined passage, but not in the individual analyses of 7a-7c.

4.3. Instrument Administration

Standard procedures were used in the administration of the GEPT-A reading test for the Taiwan-based sample. In addition, score report data from the iBT was obtained from all participants. For the U.S.-based sample, the test and survey were e-mailed to the participants as a Microsoft Word document. Participants took the test at home or in a computer lab, with the same time limit as for the regular GEPT-A. One unavoidable difference in terms of time limitations between the two groups was that the U.S. participants were given an overall time limit, while test takers in Taiwan had separate time limits for each section of the test (Careful Reading, Skimming and Scanning, and Writing). Responses were typed directly into the Word document, which was then e-mailed back to the research team. The survey was included in the same electronic document as the test.

4.4. Scoring

The tests taken in Taiwan were scored by LTTC using standard procedures. Responses to the reading tests taken in the U.S. were scored by the researchers using the key and scoring guide provided by LTTC, and it was felt that the scoring was highly consistent with those that would have been awarded by LTTC itself. The writing test essays written by U.S-based participants were scored by LTTC raters. Of the 92 essays, 86 were scored; the remaining 6 were considered non-ratable because they were off-topic or plagiarized.

4.5. Data Analysis

This section describes the steps followed in analyzing the data. It begins by describing the procedures used in the content analyses of the passages for the reading tests and integrated writing tasks, and continues with a discussion of the task analysis of the reading items. It then moves on to report the methods used to analyze the participants' responses to the reading test, writing test, and survey.

4.5.1. Content analyses of passages

Reading passages from the GEPT-A and iBT were analyzed using the Coh-Metrix Web Tool (Coh-Metrix, n.d.) and the Compleat Web VP function of the VocabProfiler (Cobb, n.d.).

Coh-Metrix is a "web-based software tool" developed at the University of Memphis to "analyze texts on multiple characteristics" (Graesser, McNamara, & Kulikowich, 2011, p. 224) and "to measure cohesion and text difficulty at various levels of language, discourse, and conceptual analysis" (Crossley, Dufty, McCarthy, & McNamara, 2007). The Coh-Metrix analysis involved the 39 variables deemed most useful from among the 106 generated by the web tool. Many of the variables excluded from this analysis were alternative measures of ones included; for example, percentile scores were reported rather than *z*-scores, on the grounds that the former were more readily interpretable.

For the vocabulary profile analysis, the "classic" word lists (K1, K2, and AWL[1]) were used, yielding four additional variables. All mid-sentence capitalized nouns automatically

[1] K1 and K2 are the 1,000 and second 1,000 most common words in English, taken from the General Service List (West, 1953). The Academic Word List (AWL; Coxhead, 2000) is a set of 570 high-frequency word families that appear in academic texts.

recategorized as K1, and any other capitalized proper nouns[2] manually recategorized as K1. Possessive forms were also recategorized when that was not done automatically.

Each passage was analyzed separately, as was the summary paragraph for GEPT R4. The summary paragraph was treated separately because while it superficially resembled a reading passage, at the same time it was essentially the combined text of six test questions. The three GEPT-A passages from the skimming task were analyzed together as GEPT R7, because they were read together as part of a single task. For both the GEPT-A and iBT, a weighted mean was computed for each variable, with the weight for a given passage based on its number of words. For the GEPT-A, the summary paragraph from Passage #4 was not included in the weighted average, as it was not deemed comparable to the other seven passages.

The same set of procedures was followed in analyzing the text of the input passages for the integrated writing tasks. For the GEPT-A, two reading passages were analyzed both as a single passage and as two separate passages. The analysis as a single passage was undertaken because both had to be read in order to attempt the writing task. The analysis as two separate passages was also performed to allow a passage-to-passage comparison with the iBT, which used a single reading passage and a single listening passage. The listening scripts for the iBT integrated writing tasks were also analyzed using the same procedures as the reading passages. Once the 43 text variables had been computed, descriptive statistics were calculated. Complete results for these variables are reported in Appendix A for GEPT-A reading passages, iBT reading passages, GEPT-A (reading) input passages for the integrated writing task, and iBT reading and listening input passages for the integrated writing tasks.

4.5.2. Analysis of participant test performance data

Descriptive statistics. Descriptive statistics were computed for participant demographic information, total GEPT reading and writing scores, scores on both sections of the GEPT, and iBT scores for reading, writing, listening, and speaking. Descriptive statistics were also computed for each of the six items on the survey.

Item and reliability analyses were performed on the GEPT-A reading test. In the absence of appropriate cut scores, point-biserial coefficients were used to estimate item discrimination.[3] The discrimination was calculated using adjusted total scores—that is, with the item in question being removed from the total in order to avoid inflation of the coefficient due to autocorrelation.

In addition, GEPT-A reading and writing task 1 scores were regressed on iBT reading and writing scores.

Correlations and exploratory and confirmatory factor analyses. Correlations were calculated among the two GEPT-A reading sections, the GEPT-A integrated writing task, the iBT sub-

[2] This excluded uncapitalized words that were part of proper nouns, such as the Italian words *di* and *dei* in GEPT R1. This choice was made because a non-native speaker reading the passages and not familiar with a particular third language could be expected to recognize any capitalized word as a proper noun, but could not necessarily be expected to recognize that the foreign word was part of a longer, multiword proper noun. This decision not to recategorize such words recognizes the additional level of vocabulary knowledge and/or top-down reading ability necessary to comprehend these words.

[3] In the case of polytomous items, this was technically a Pearson *r*, but since point-biserials are Pearsonian correlation coefficients, the two coefficients are essentially the same.

scores, and responses to the six survey questions, resulting in a 13 x 13 correlation matrix. Pearson *r* was used because with a very few minor exceptions, the variables had relatively normal distributions (i.e., means and medians close together, and skewedness and kurtosis with absolute values of less than 2). Two-tailed significance tests were used; while it was assumed that the seven test score variables would have positive relationships, and that there might be a negative relationship between perceptions of GEPT-A difficulty and performance on the GEPT-A and iBT, any relationships between the "relevance" survey questions and the other variables could not be predicted in advance.

An exploratory factor analysis was conducted using SPSS of the GEPT-A reading and writing scores and the iBT reading and writing scores. For the GEPT-A reading, items were grouped into seven testlets, each based on one of the reading passages. Testlet scores were used rather than scores on individual items, because item-level scores often lack sufficient variance—particularly in smaller datasets such as this one—for clear factor structures to emerge. The factors were extracting using principal axis factoring. Initially, no minimum criterion or maximum number of factors to extract was set. After the initial extraction, any factors with no unrotated loadings under .30 were dropped (following Comrey, 1992), and the analysis was repeated. When the model would not converge for a given number of factors, extraction was attempted with one fewer. When the number of factors stabilized, the result was rotated using the Varimax algorithm. Any factor that did not have at least three variables with loadings of .30 or greater was dropped, and the procedure was repeated. If a model would not converge, the analysis was attempted with one factor fewer. Once the factor structure had stabilized following this procedure, a solution with one additional factor[4] was tried for comparison. These models were compared on the basis of simple structure, parsimony, and interpretability in order to determine the number of factors in the final model. This model was then rotated using Promax, which yielded an oblique factor structure.

A confirmatory factor analysis was then conducted using AMOS, taking the results of the exploratory factor analysis as a starting point (Model 1). The steps for CFA followed standard procedures outlined in Kunnan (1998). For purposes of comparison, additional CFAs were performed with all GEPT-A reading passages and iBT reading scores loading on one factor, and both GEPT-A and iBT writing scores loading on a second factor (Model 2). The two factors were set to correlate with each other, on the assumption that reading and writing are related aspects of language ability. A third model (Model 3) was also tested—primarily for the purposes of exclusion—with the GEPT variables loading on one factor and the iBT reading and writing scores loading on a second factor (which was correlated with the first), to test the hypothesis that the two tests measure separate but related things. Additional models were also tested in an attempt to achieve satisfactory model fit.

Goodness of fit in the CFA models was evaluated using χ^2, the NFI, NNFI, CFI, and RMSEA. Consideration was also given to the significance of parameter estimates, as tested using the ratio of raw parameter estimates to their standard errors (Byrne, 2010). All fit indices were provided by AMOS, with the exception of the NNFI, which was calculated by hand from AMOS output following Hu and Bentler (1995). We employed a variety of fit indices in order to evaluate fit from multiple perspectives. The least emphasis was placed on χ^2, since it

[4] Normally, a solution would have been attempted with one factor fewer, too, but that proved not to be possible in this case, since the final model only had one factor.

almost invariably is significant for any model tested, despite whatever other fit indices may show.

5. Results

In this section, we present the findings of the study. We begin with background information on the study participants, followed by the results of the content analyses of the passages. We then continue with the task analysis of the reading items, and conclude with the analysis of the test and survey results.

5.1. Participant Background Information

Tables 2 to 7 summarize the background information of the participants in the study. Table 2 summarizes the breakdowns of ages and genders for all participants in the sample, and Tables 3 and 4 detail their academic status (undergraduate or postgraduate) and academic majors. Table 5 summarizes when participants took the iBT. Tables 6 and 7 then provide information on the first language backgrounds of the U.S.-based participants, and when they arrived in the United States.

Table 2. Participants' Ages and Genders

Age	n	Percentage	Gender	n	Percentage
17-24	83	45.4%	Male	82	44.6%
25-30	85	46.4%	Female	102	55.4%
31 and above	15	8.2%			
Total	183	100.0%		184	100.0%

Table 3. Participants' Academic Status

	n	Percentage
Undergraduate	70	38.0%
Postgraduate	114	62.0%
Total	184	100.0%

Table 4. Participants' Academic Majors

	n	Percentage
Agriculture	4	2.2%
Arts	6	3.3%
Business	32	17.4%
Education	13	7.1%
Engineering & computer science	63	34.2%
Health professions	7	3.8%
Humanities	36	19.6%
IEP	2	1.1%
Math & science	3	1.6%
Other or undeclared	4	2.2%
Social Sciences	14	7.6%
Total	184	100.0%

Table 5. Participants' iBT Test Dates

Test year	*n*	Percentage
Before 2012	11	6.0%
2012	18	9.8%
2013	120	65.2%
2014	34	18.5%
2015	1	0.5%
Total	184	100.0%

Note. GEPT tests were all taken in 2014 or 2015.

Table 6. First Language Backgrounds of U.S.-Based Participants

Language	*n*	Percentage
Arabic	2	2.2%
Bengali	1	1.1%
Chinese	26	28.3%
Dari & Pashto	1	1.1%
Farsi	2	2.2%
German	1	1.1%
Hindi	5	5.4%
Indonesian	1	1.1%
Japanese	1	1.1%
Kannada	6	6.5%
Korean	14	15.2%
Marathi	7	7.6%
Portuguese	1	1.1%
Portuguese & French	1	1.1%
Punjabi	1	1.1%
Russian	1	1.1%
Spanish	3	3.3%
Tamil	1	1.1%
Telugu	8	8.7%
Telugu & Hindi	1	1.1%
Thai	2	2.2%
Urdu	2	2.2%
Vietnamese	4	4.3%
Total	92	100.0%

Table 7. Participants' Time of Arrival in U.S.

Year	*n*	Percentage
Before 2012	13	14.6%
2012	4	4.5%
2013	9	10.1%
2014	59	66.3%
2015	4	4.5%
Total	89	100.0%

5.2. Content Analysis of Passages

The Coh-Metrix and LexTutor analyses of the reading passages resulted in 43 values for each reading passage. These results are presented in Appendix A. The independent samples Mann-Whitney U test performed on the 43 variables indicated that only six variables were significantly different across the two tests. As shown in Table 8, the two tests varied significantly in the number of words per passage, two separate measures of lexical diversity (MTLD and VOCD) across all words, and in terms of mean number of modifiers per noun phrase, mean sentence syntax similarity across paragraphs, and the percentage of K1 words in each passage. The SD (calculated without weighting for passage length) for the variables with significant differences is also presented in Table 8 as a measure of effect size.[5]

Table 8. Significant Differences Between the GEPT-A and iBT Reading Passages

Variable	GEPT Mean [a]	iBT Mean [a]	SD [b]	*p*
DESWC (Word count, number of words)	736.8	687.4	82.3	.023
LDMTLDa (Lexical diversity, MTLD, all words)	106.01	84.54	22.38	.023
LDVOCDa (Lexical diversity, VOCD, all words)	106.90	86.71	16.00	.005
SYNNP (Number of modifiers per noun phrase, mean)	0.89	1.06	.15	.023
SYNSTRUTt (Sentence syntax similarity, all combinations, across paragraphs, mean)	.10	.08	.01	.000
K1	79.69	75.51	3.82	.023

[a] Weighted by number of words per passage.

[b] Unweighted, for all GEPT-A and iBT passages combined.

The Coh-Metrix and LexTutor analyses of the input passages for the integrated writing tasks (reading only for the GEPT-A; reading and listening for the iBT) also resulted in 43 values for each passage. These results are presented in Appendix A. The results of the independent samples Kruskal-Wallis test were not significant for any of the text variables. This may have been because of the very small sample size (*n* = 2, 3, and 3 for the three groups of passages), although it is not possible to say so definitively.

5.3. Task Analysis of Reading Items

This section reports on the findings of the task analysis of the reading sections of the GEPT-A and iBT. It begins by reporting the topical content of the various reading passages, and then

[5] More conventional measures of effect size were not available because SPSS does not provide *U*, just the significance of the test statistic.

examines the aspects of the reading construct that individual items seemed most likely to assess, the scope of the items, and the task formats used on the two tests.

The listing of passages and their topics is contained in Table 9. As can be seen from some of the classifications, some passages were more challenging than others to assign to a single subject matter category. The GEPT-A passages came from a range of subject matter topics, but with no content from the life sciences. In contrast, the iBT passages covered the same sorts of topics as the GEPT, but with the addition of life sciences topics as well. Notably, the iBT physical sciences passages all dealt with geology.

Table 9. Titles and Topics of the Reading Comprehension Passages

Passage	Title [a]	Topic
GEPT R1	Caravaggio	Art history
GEPT R2	Value-Added Teacher Ratings	Education
GEPT R3	Hydrates	Physical sciences (geology)
GEPT R4	Brownfield Redevelopment	Social sciences (economics, sociology)
GEPT R5	Hudson's Bay Company	Social sciences (history)
GEPT R6	Victor the Wild Child	History of science
GEPT R7	Three Historical Attractions	Tourism
iBT R1	Nineteenth-Century Politics in the United States	Social sciences (history)
iBT R2	The Expression of Emotions	Psychology
iBT R3	Geology and Landscape	Physical sciences (geology)
iBT R4	Feeding Habits of East African Herbivores	Life sciences (biology)
iBT R5	Loie Fuller	Art history (performing arts—dance)
iBT R6	Green Icebergs	Physical sciences (geology?)
iBT R7	Architecture	Architecture (art history?)
iBT R8	The Long-Term Stability of Ecosystems	Life sciences (biology)
iBT R9	Depletion of the Ogallala Aquifer	Physical sciences (geology)

[a] Titles for Part 1 were not included on the test form, and are taken from the GEPT-A Marking Scheme. Titles for the first two passages in Part 2 are taken from the test. The third passage was a collection of three separately-titled shorter passages, and the title of the overall whole was inferred by the researchers.

Complete lists of the item-by-item findings are presented in Appendix B, but the results are summarized for the GEPT-A and iBT in Table 10 and Figure 1. One issue that presented itself involved vocabulary questions—whether they were tapping into the top-down reading process of identifying the meaning of unfamiliar vocabulary from context clues, or were instead assessing vocabulary knowledge, a point which is treated in the Discussion below. In rating item scope, however, it was assumed that they are in fact assessing the top-down reading process, not knowledge of vocabulary.

Table 10. Summary of Construct Coverage for the GEPT-A and iBT

Construct component	# of GEPT items	% of GEPT items	# of iBT items	% of iBT items
Reading for specific details	11	27.5%	42	34.4%
Reading for the main idea	1	2.5%	0	0.0%
Reading for major points	0	0.0%	8	6.6%
Inferencing	0[a]	0.0%	8	6.6%
Identifying author purpose	1	2.5%	9	7.4%
Vocabulary knowledge/ determining the meaning of unfamiliar vocabulary from context	0	0.0%	31	25.4%
Vocabulary knowledge	0	0.0%	5	4.1%
Sensitivity to rhetorical organization	1	2.5%	9	7.4%
Sensitivity to cohesion	0	0.0%	2	1.6%
Paraphrasing and/or summarizing	6	15.0%	8	6.6%
Skimming	12	30.0%	0	0.0%
Scanning	8	20.0%	0	0.0%
Total	40	100.0%	122	100.0%

[a] LTTC considers six items (15.0% of all items) to be inference questions; four of those items (10.0%) are classified here as reading for specific details, and two (5.0%) are counted as paraphrasing and/or summarizing.

As can be seen, the construct coverage of the two tests is similar in that they both have more items requiring reading for specific details than any other part of the reading construct. Both include paraphrasing and/or summarizing, although the GEPT-A assesses this more extensively than does the iBT. Neither does much to assess the ability to identify the main idea of a passage (the GEPT-A included one such item in its careful reading section; the scanning section primarily requires scanning for the main idea of a paragraph, as opposed to the main idea of an entire passage), although the iBT does include a number of items that appear to assess the ability to read for major points or ideas—one for nearly every passage.

The tests also differ markedly in several ways. The GEPT-A devotes heavy coverage to skimming and scanning, something the iBT ignores. In contrast, the iBT includes a large number of items assessing vocabulary knowledge or the ability to determine the meaning of unfamiliar vocabulary from context. Finally, another major difference between the two tests lies in the areas of top-down reading processes such as inferencing, identifying author purpose, and sensitivity to rhetorical organization and cohesion. The iBT includes these to a far greater extent than does the GEPT-A.

Table 11 and Figure 2 describe the breakdown of the scope of the reading items on the GEPT-A and iBT. As can be seen, the iBT predominantly uses items with a narrow or very narrow scope (i.e., requiring the processing of several sentences or less). The GEPT-A, on the other hand, focuses more on moderate-scope items (i.e., those requiring the processing of an entire paragraph (or close to it), with this level of scope the proving to be the most common one. Finally, the iBT included a high proportion of items with broad or very broad scope (i.e., the key information was spread across multiple paragraphs or the entire passage, respectively). The GEPT-A, on the other hand, had a much lower proportion of items with these levels of scope, with 8 of the 11 coming from the scanning section.

Table 11. Summary of Scope of Reading Items for the GEPT-A and iBT

Item scope	# of GEPT items	% of GEPT items	# of iBT items	% of iBT items
Very narrow	2	5.0%	41	33.6%
Narrow	8	20.0%	51	41.8%
Moderate	17	42.5%	14	11.5%
Broad	3	7.5%	5	4.1%
Very broad	10	25.0%	11	9.0%
Total	40	100.0%	122	100.0%

Table 12. Summary of Task Formats of Reading Items on the GEPT-A and iBT

Task format	# of GEPT items	% of GEPT items	# of iBT items	% of iBT Items
Short answer	15	37.5%	0	0.0%
Multiple choice	5	12.5%	113	92.6%
Multiple response multiple choice	0	0.0%	8	6.6%
Fixed multiple choice	8	20.0%	0	0.0%
Matching	12	30.0%	0	0.0%
Categorization	0	0.0%	1	.8%
Total	40	100.0%	122	100.0%

In the last portion of the reading test task analysis, we compared the task formats used on the two tests. The results for this are summarized in Table 12 and Figure 3. While the iBT was entirely dependent upon selected response items, the GEPT-A included a substantial proportion of short answer questions, with only about a third of the items using traditional multiple choice. In contrast, the iBT mainly relied upon multiple choice items, with only 8% of the total items representing task formats *other* than multiple-choice.

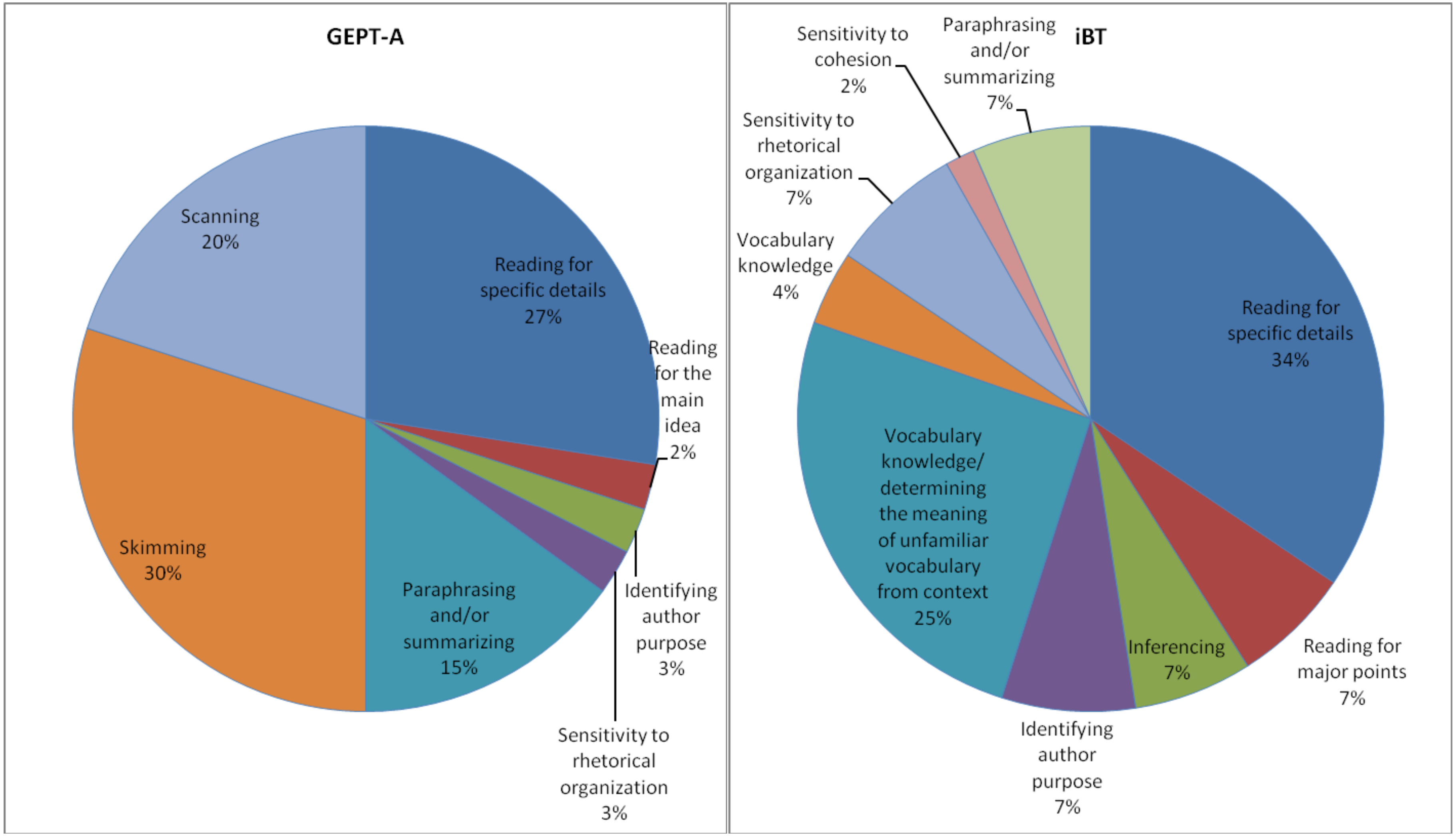

Figure 1. Reading Construct Coverage for the GEPT-A and iBT

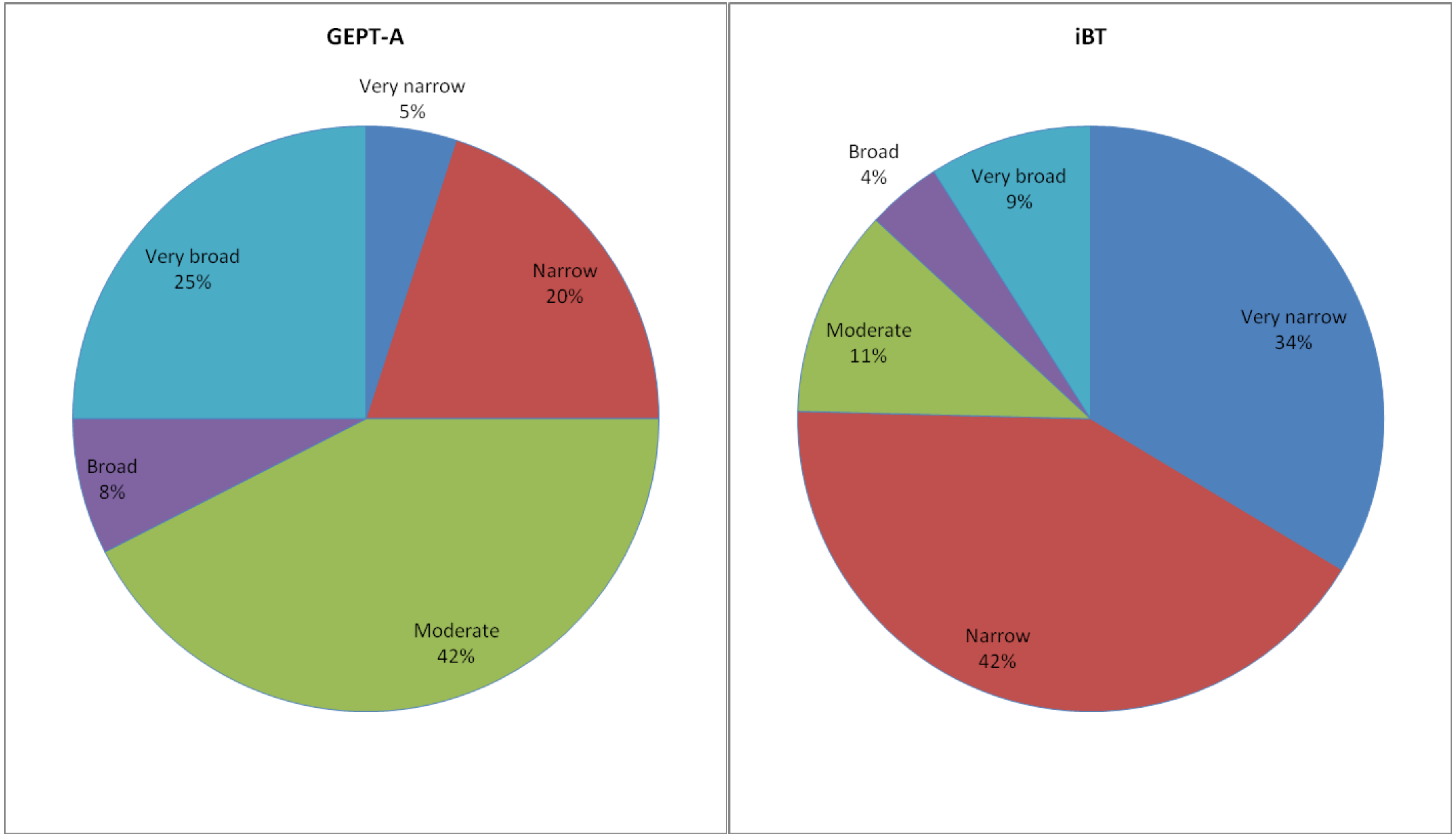

Figure 2. Scope of Reading Items on the GEPT-A and iBT

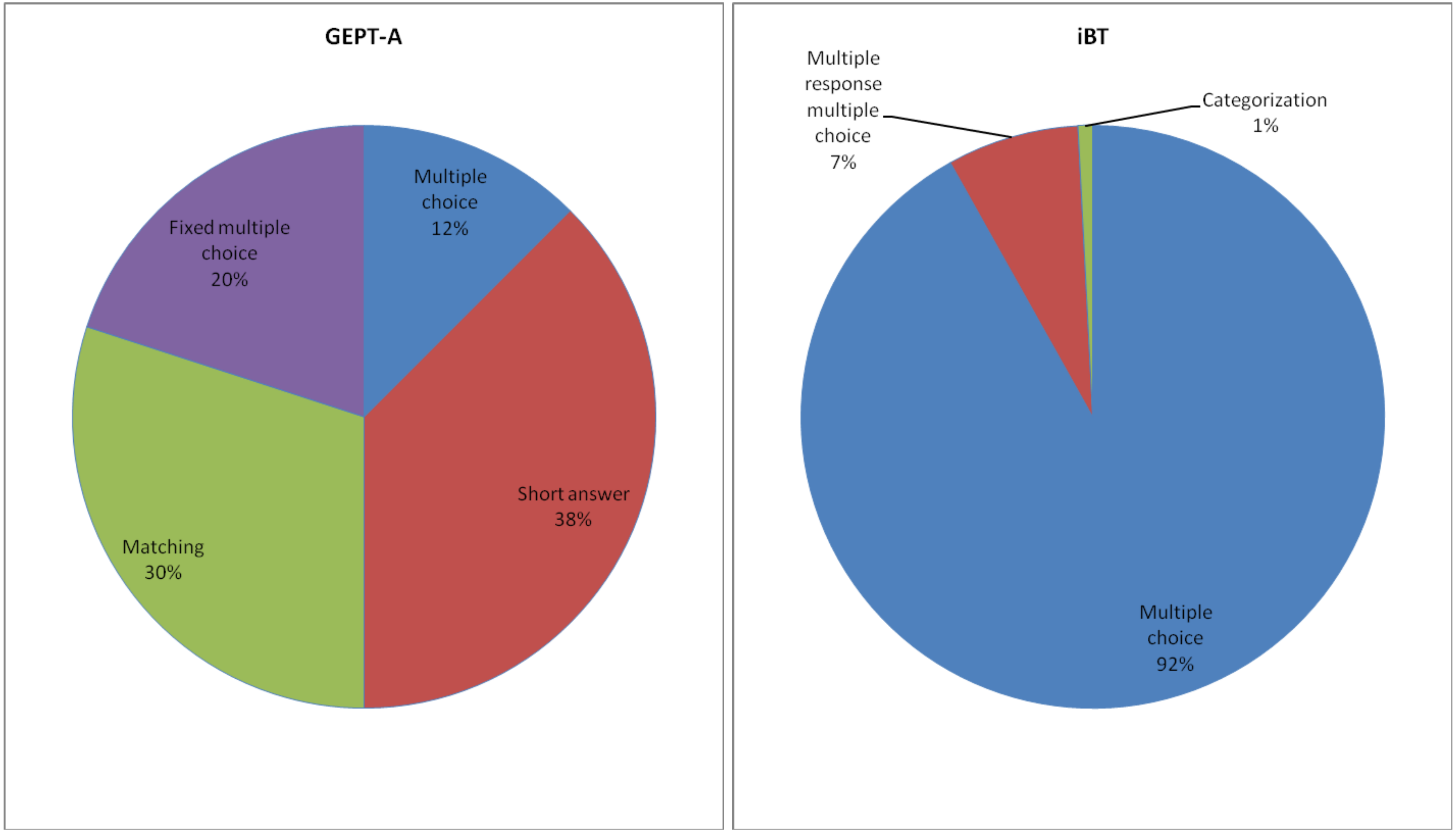

Figure 3. Task Formats of Reading Items on the GEPT-A and iBT

5.4. Analysis of Participant Test Performance Data

Table 13 provides the descriptive statistics for GEPT-A and iBT scores. Scores are reported in percentages[6] for comparability; raw scores, including Cronbach's alpha and the standard error of measurement (SEM) for the GEPT-A Reading test, are provided in Appendix C, Table C1. Similarly, descriptive statistics for total score by passage are presented in Table 14 for percentages, while the descriptives for raw scores are provided in Table C2.

Table 13. Descriptive Statistics for GEPT and iBT Scores (Percentage Scores)

	GEPT Reading 1	GEPT Reading 2	GEPT Reading	GEPT Writing	iBT Reading	iBT Writing	iBT Listening	iBT Speaking
Mean	57.5%	58.2%	57.9%	51.1%	82.9%	80.1%	81.2%	74.0%
Median	57.5%	60.0%	58.8%	50.0%	86.7%	83.3%	83.3%	76.7%
SD	16.9%	22.7%	18.3%	9.9%	14.0%	11.6%	14.3%	10.1%
Q	10.3%	17.5%	13.8%	4.9%	8.3%	8.3%	9.2%	6.7%
Skewness	-0.3	-0.2	-0.2	1.3	-1.9	-1.0	-1.5	-0.4
Kurtosis	0.2	-0.9	-0.6	1.6	5.4	1.7	3.2	1.2
Alpha	0.774	0.818	0.880	--	--	--	--	--
SEM	8.0%	9.7%	6.3%	--	--	--	--	--

Table 14. Descriptive Statistics for Percentage Scores on Individual GEPT-A Passages (all test takers)

	R1	R2	R3	R4	R5	R6	R7
Mean	63.7%	69.3%	62.1%	39.8%	45.3%	69.7%	59.2%
Median	62.5%	70.0%	60.0%	41.7%	33.3%	83.3%	62.5%
SD	22.4%	19.0%	24.4%	26.9%	29.0%	31.0%	29.3%
Q	12.5%	10.0%	15.0%	20.8%	25.0%	25.0%	25.0%
Skewness	-0.4	-0.6	-0.6	0.1	0.3	-0.7	-0.4
Kurtosis	-0.4	0.3	0.0	-0.9	-0.9	-0.7	-0.7

[6] In the case of the four iBT scores, as well as the GEPT writing score and overall GEPT reading score, the scores are not percentage correct, but percentage of the maximum possible number of scale points. For GEPT Reading 1 and Reading 2, they *are* the percentage correct. GEPT Reading = Reading 1 x 1.5 + Reading 2 x 3.

Table 15 presents the item analysis results for the GEPT-A reading items. Discrimination was calculated based on adjusted total score for the particular section of the reading test (i.e., with each item's score subtracted from the total, to prevent autocorrelation). Tables C3 and C4 include item analysis results based on total GEPT score and on total passage scores as well (also adjusted for autocorrelation).

Table 15. Item Analysis Results for GEPT-A Reading

Item	IF*	Discrimination	Item	IF*	Discrimination
R01	0.80	0.35	R21	0.40	0.30
R02	0.58	0.39	R22	0.68	0.30
R03	0.61	0.25	R23	0.33	0.15
R04	0.57	0.27	R24	0.49	0.40
R05	0.88	0.23	R25	0.42	0.32
R06	0.64	0.32	R26	0.40	0.54
R07	0.43	0.26	R27	0.81	0.43
R08	0.91	0.33	R28	0.66	0.39
R09	0.61	0.21	R29	0.70	0.42
R10	0.70	0.31	R30	0.70	0.40
R11	0.51	0.35	R31	0.65	0.39
R12	0.54	0.36	R32	0.66	0.47
R13	0.54	0.41	R33	0.71	0.37
R14	0.81	0.30	R34	0.61	0.31
R15	0.46	0.31	R35	0.50	0.37
R16	0.17	0.34	R36	0.57	0.42
R17	0.31	0.51	R37	0.64	0.33
R18	0.60	0.36	R38	0.57	0.38
R19	0.58	0.39	R39	0.55	0.50
R20	0.27	0.48	R40	0.60	0.56

Note. Discrimination was calculated as the correlation between items and total score on that section of the test, adjusted for autocorrelation.

The results of the test takers survey are summarized in Tables 16 to 18, which present descriptive statistics, response frequencies for the two questions on level of difficulty, and response frequencies for the questions dealing with relevance to academic studies, respectively.

Table 16. Descriptive Statistics for the Test Takers Survey

	Difficulty level		Content relevant to academic studies		Tasks relevant to academic studies	
	Reading	Writing	Reading	Writing	Reading	Writing
Mean	2.2	1.9	3.1	2.7	3.2	2.5
Median	2	2	3	2	3	2
SD	0.6	0.6	1.0	1.0	1.2	1.2
Q	0.5	0.0	1.0	1.0	1.0	0.5
Skewness	-0.1	-0.1	-0.1	0.3	-0.1	0.6
Kurtosis	-0.5	0.2	-0.9	-0.8	-1.0	-0.7

Note. The two questions on difficulty level were rated on a 1-3 scale (*easy*, *medium*, *difficult*); the four questions on content and task relevance were rated on a 1-5 scale (*strongly agree*, *agree*, *neither agree nor disagree*, *disagree*, *strongly disagree*).

Table 17. Response Frequencies for the Test Takers Survey: Level of Difficulty Questions

	Reading		Writing	
	n	%	*n*	%
Easy	17	9.7%	36	20.5%
Medium	103	58.5%	121	68.8%
Difficult	56	31.8%	19	10.8%
Total	176	100.0%	176	100.0%

Table 18. Response Frequencies for the Test Takers Survey: Questions on Relevance to Academic Studies

	Content relevant				Tasks relevant			
	Reading		Writing		Reading		Writing	
Response	*n*	%	*n*	%	*n*	%	*n*	%
Strongly agree	8	4.6%	17	9.8%	11	6.3%	34	19.5%
Agree	53	30.3%	73	42.0%	48	27.6%	75	43.1%
Neither agree nor disagree	43	24.6%	39	22.4%	36	20.7%	24	13.8%
Disagree	60	34.3%	39	22.4%	57	32.8%	31	17.8%
Strongly disagree	11	6.3%	6	3.4%	22	12.6%	10	5.7%
Total	175	100.0%	174	100.0%	174	100.0%	174	100.0%

The results of the regression analysis for the GEPT-A reading scores was the following equation: GEPT-A-R= 8.561 + 2.458* iBT-R (R^2= 0.218, SEE=19.54. The results of the regression analysis for the GEPT-A writing task 1 score was GEPT-A-W1= 1.225 + 0.055* iBT-W (R^2 = 0.148, SEE=0.46.

5.4.1. Analyses regarding relationships among tests

In this section, we present the results of the three stages of correlational analyses that were performed. We begin with the correlations among key variables, move on to exploratory factor analyses of the GEPT-A and iBT reading and writing scores, and conclude with confirmatory factor analyses of those scores.

Correlations among variables. Table 19 contains the correlation matrix for the two GEPT-A reading sections, GEPT-A first writing task, all four iBT subscores, and responses to the six test takers survey questions. Table D1 repeats the matrix with the actual significance level and sample size[7] for each correlation. Unsurprisingly, all of the correlations among test scores were highly significant ($p \leq .001$) for the variables associated with GEPT-A and iBT scores. The GEPT-A reading and iBT reading scores were correlated at $r = .467$. While a medium-to-large correlation, this indicates only about 22% shared variance between the two tests. Similarly, the medium correlation ($r = .385$) between the GEPT and iBT writing scores indicates about 15% shared variance. The correlation between the two GEPT-A reading sections was very high, with an effect size of $r^2 = .491$. The effect sizes for the GEPT-A reading and writing scores were somewhat smaller ($r^2 = .283$ and .168, respectively). The correlations among the four iBT scores had smaller effect sizes, ranging between 10.8% and 38.4% shared variance for each pair of variables. Correlations across the GEPT-A and iBT variables were somewhat lower overall than those among the scores from within a given test. Effect sizes ranged from near-trivial ($r^2 = .073$) to modest but appreciable ($r^2 = .212$).

As anticipated, there was a negative relationship between test scores and perceptions of GEPT-A difficulty, but it was not significant for every test score variable, and had a minor effect size at most. The largest correlation was between perceptions of GEPT-A reading test difficulty and GEPT-A Reading Task 1 scores—a significant relationship with a minor effect size ($r^2 = .147$). Perhaps the most interesting of the correlations with perceived difficulty was the one between the perception of difficulty for the GEPT-A reading and GEPT-A writing tests, which had 11.4% shared variance—a minor relationship, yet nevertheless the second-largest in this set of correlations. Also worth noting was the low correlation between perception of GEPT-A writing difficulty and GEPT-A writing score ($r^2 = .031$).

The relationship between the perceived relevance of GEPT-A content (to participants' academic studies) and other variables was for the most part not significant. There was only one significant relationship between content relevance and test scores on the GEPT-A or iBT: the correlation between GEPT-A writing scores and the perceived relevance of the content of the GEPT-A *reading* test. Given its trivial effect size ($r^2 = .037$), it may have been a spurious correlation (i.e., resulting from chance). Content relevance of the reading and writing tests were closely related, with strong effect size ($r^2 = .419$). There was a significant but trivial relationship between the content relevance of the GEPT-A writing test and participants' perception of the difficulty of the GEPT-A writing tasks, as well as a significant but trivial

[7] There was some minor variation in sample size across correlations because of missing data—some participants left certain questions unanswered. Pairwise deletion of cases was used in computing the correlations.

negative relationship between GEPT-A writing content relevance and GEPT-A reading difficulty.

The relationship between the perceived relevance of GEPT-A tasks (to participants' academic studies) and other variables was minor, and difficult to interpret. There were small significant negative correlations with GEPT-A reading and writing scores, except for the non-significant correlation between reading task relevance and GEPT writing scores. None of these correlations had effect sizes greater than 6.7% shared variance. These correlations mean that as participants' perception of task relevance increased, their scores went down, and vice versa. There was a small significant correlation between writing task relevance and perceived writing task difficulty, but similarly, the effect size (r^2 = .058) bordered on trivial. There was a small significant but borderline-trivial correlation between perceived GEPT-A writing difficulty and the relevance of its writing tasks to participants' academic studies, indicating that—to a small extent—participants tended to associate GEPT-A writing difficulty with relevance of the writing tasks to their own academic studies. There were also significant correlations between perceptions of GEPT content relevance and task relevance for both reading and writing, with the highest occurring between writing task relevance and writing content relevance, and between reading task relevance and writing task relevance. These last two correlations both showed medium effect sizes (r^2 = .283 and .231, respectively).

Exploratory factor analysis. The results of the EFA are summarized in Table 20. Correlations among the variables analyzed are presented in Table D2. A single-factor solution provided relatively high loadings for all 10 variables, and was both parsimonious and very easy to interpret.

In contrast, a correlated two-factor solution (r = .593) yielded a first factor that accounted for most of the GEPT-A reading passages (Passages 3-7), and a second factor on which the iBT reading and writing scores loaded (with reading particularly high). The GEPT-A writing and the remaining two GEPT-A reading passages cross-loaded on both factors, with roughly equal loadings on each. A three-factor solution would not converge. More than three factors would have led to model identification problems, since it would have required at least one factor with only two indicator variables. Therefore, the single-factor model was confirmed as the best one in the EFA.

Table 19. Correlations among GEPT Section Scores, iBT Section Scores, and Test Takers Survey Responses

	GEPT_R1	GEPT_R2	GEPT_W	iBT_R	iBT_W	iBT_L	iBT_S	DifficR	DifficW	CntRelR	CntRelW	TskRelR	TskRelW
GEPT_R1	1.000												
GEPT_R2	.701**	1.000											
GEPT_W	.532**	.410**	1.000										
iBT_R	.457**	.414**	.338**	1.000									
iBT_W	.425**	.316**	.385**	.545**	1.000								
iBT_L	.460**	.332**	.317**	.620**	.591**	1.000							
iBT_S	.388**	.270**	.359**	.328**	.603**	.552**	1.000						
DifficR	-.384**	-.336**	-.216**	-.181*	-.342**	-.190*	-.279**	1.000					
DifficW	-.176*	-.052	-.177*	-.197**	-.223**	-.203**	-.197**	.337**	1.000				
CntRelR	-.091	-.064	-.193*	-.024	-.105	-.070	-.063	.101	.125	1.000			
CntRelW	.146	.144	-.077	.000	-.032	-.067	.014	-.160*	.186*	.647**	1.000		
TskRelR	-.151*	-.259**	-.091	.046	-.114	.083	-.041	.130	-.007	.304**	.177*	1.000	
TskRelW	-.156*	-.168*	-.218**	-.116	-.079	-.127	-.048	.010	.241**	.373**	.532**	.481**	1.000

**Correlation is significant at the 0.01 level (2-tailed).

*Correlation is significant at the 0.05 level (2-tailed).

Table 20. EFA Results for GEPT-A reading and writing and iBT reading and writing scores

Variables	Factor Loadings
GEPT_Pssg1	.628
GEPT_Pssg2	.563
GEPT_Pssg3	.616
GEPT_Pssg4	.597
GEPT_Pssg5	.557
GEPT_Pssg6	.576
GEPT_Pssg7	.615
GEPT_W	.609
iBT_R	.595
iBT_W	.547

Confirmatory factor analysis. As explained previously, the first model tested took the results of the EFA as its starting point—a single-factor solution, with the factor assumed to represent reading and writing ability. All parameter estimates were significant, and the model converged in eight iterations. The path diagram for this model is presented in Figure 4.

Model 2, which featured all reading scores loading on one factor and both writing scores loading on a second factor, was attempted next. All parameter estimates were significant, and the model converged in eight iterations. However, as Table 21 indicates, fit was largely unaffected. Because iBT reading and writing scores were self-reported data, and were the only variables in this analysis for which there were missing data, the error terms for these two variables were correlated. This resulted in a noticeable improvement in model fit. The path diagram of the resulting model is shown in Figure 5. The fit borders between moderate and mediocre, as none of the fit indices reach the cutoff values for "good" model fit. Nevertheless, since this model had the best fit, and was also the best in terms of interpretability, it was determined to be the best possible with the current data.

A third model (Model 3) was also tested, with the GEPT variables loading on one factor and the iBT reading and listening scores loading on a second factor (which was correlated with the first), to test the hypothesis that the two tests measure separate but related things (see Figure 6). As with the other two models, the model converged in eight iterations, and all parameter estimates were significant. Model fit was slightly better than with the unmodified Model 1 and Model 2, but was still mediocre at best (see Table 21). Since this model was only identified if missing data were imputed—unlike Models 1 and 2—it was decided that Model 3 was not an accurate description of the factor structure being analyzed.

Table 21 summarizes the fit of all the models tested; as can be seen, this model fit the data rather poorly.

A few additional models were attempted with task factors or cross-loadings on multiple factors for selected observed variables, in hopes of finding a model with better fit. These models tended not to converge at all, and when they did, always resulted in worsened fit. Model 2 was thus confirmed as the best available model that included both the GEPT-A and IBT under the circumstances. Additional EFA analyses were then conducted separately on GEPT-A scores and iBT scores. The GEPT-A was modeled with writing scores included in one model (Model 4) and excluded in another (Model 5). To help model identification in the iBT analysis, listening and speaking scores were included along with the reading and writing. In both exploratory analyses, a single-factor solution proved to be the most interpretable result. The factor matrices for these are presented in Tables 22 and 23, and the goodness of fit summaries for the models are presented in Table 24. Figures 7 and 8 show the path diagrams for the three models. Model 5 failed to converge, and AMOS indicated it was probably underidentified.[8] Since model fit was improved but still mediocre for Model 5, an additional CFA was attempted using only GEPT-A reading data. The resulting Model 6 (see Table 24 and Figure 9) had the best fit of any model in the analyses. It had a non-significant χ^2, and the NNFI and CFI were both consistent with good fit. The NFI and RMSEA were not as good; taken together, this suggests that the model had marginally good fit.

Table 21. Goodness of Fit Summary for Models

Statistic	Model 1	Model 2	Model 2a	Model 3
χ^2	92.085	90.567	62.940	66.731
df	35	34	33	34
p	.000	.000	.001	.001
NFI	.829	.832	.883	.876
NNFI	.815	.811	.897	.891
CFI	.882	.883	.938	.932
RMSEA	.094	.095	.070	.073
RMSEA CI$_{.90}$	.071 - .118	.072 - .119	.043 - .097	.046 - .098

Note. Model 1 = 1-Factor reading and writing; Model 2 = 2-Factor reading & writing
Model 2a = 2-Factor reading & writing with correlated errors; Model 3 = 2-factor GEPT-A & iBT

[8] As there is no obvious reason why this model should have been unidentified, it may be a case of empirical under-identification (Rindskopf, 1984). This point is addressed further in the discussion section.

Table 22. EFA Results for GEPT-A Reading & Writing Scores

Variables	Factor Loadings
GEPT_Pssg1	.608
GEPT_Pssg2	.548
GEPT_Pssg3	.639
GEPT_Pssg4	.627
GEPT_Pssg5	.553
GEPT_Pssg6	.607
GEPT_Pssg7	.646
GEPT_W	.602

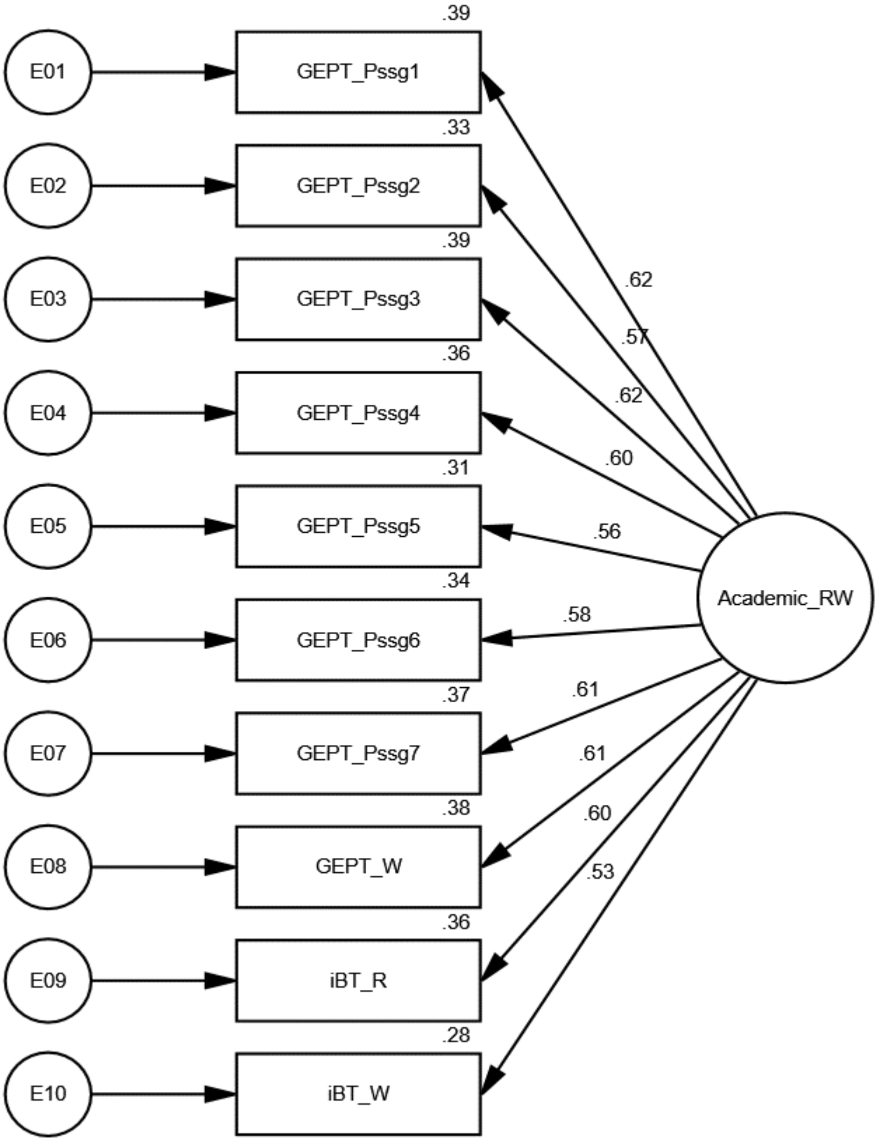

Figure 4. Path Diagram for Model 1(1-Factor Reading), with Parameter Estimates

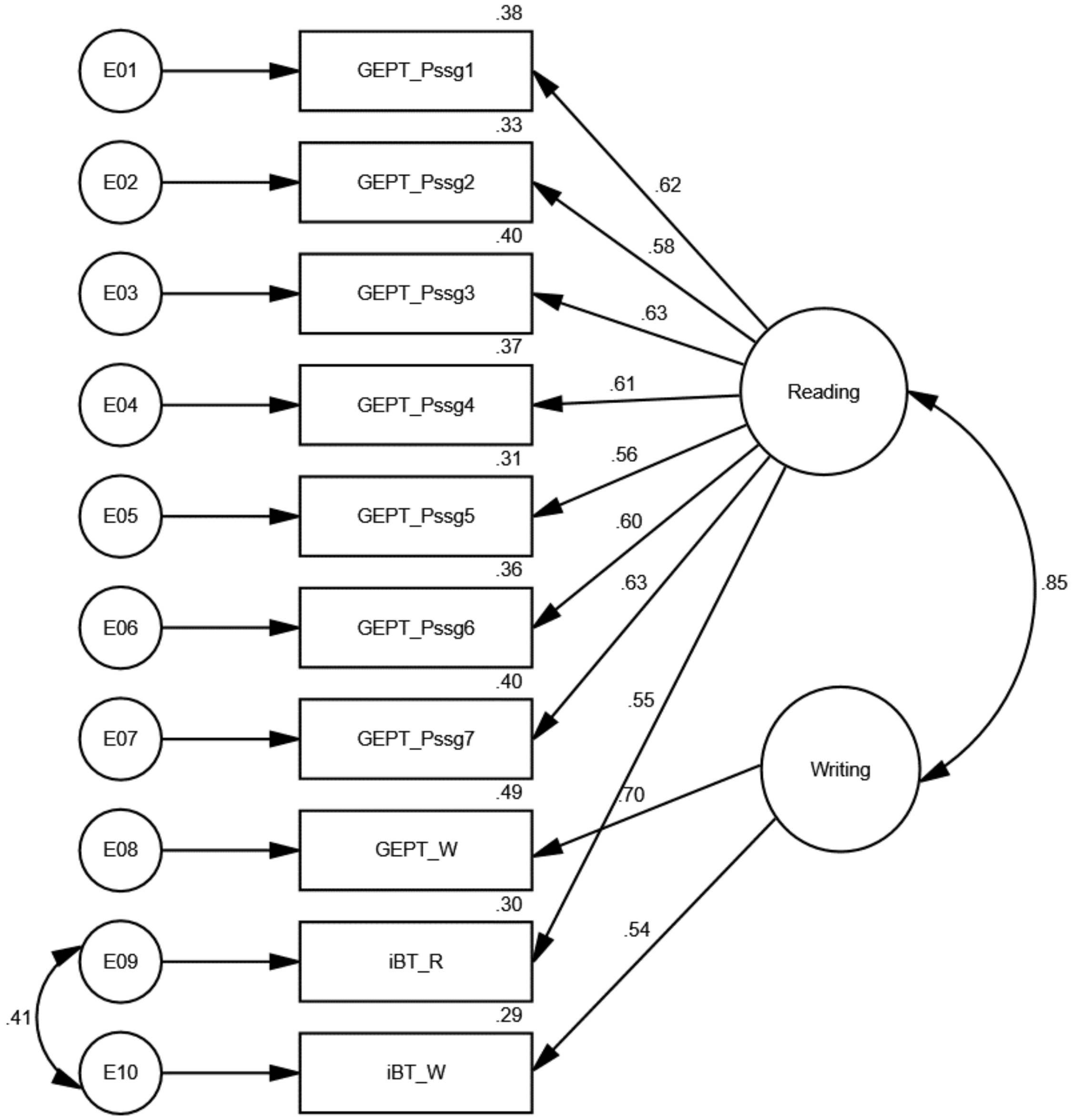

Figure 5. Path Diagram for Model 2a (2-Factor Reading & Writing with correlated errors), with Parameter Estimates

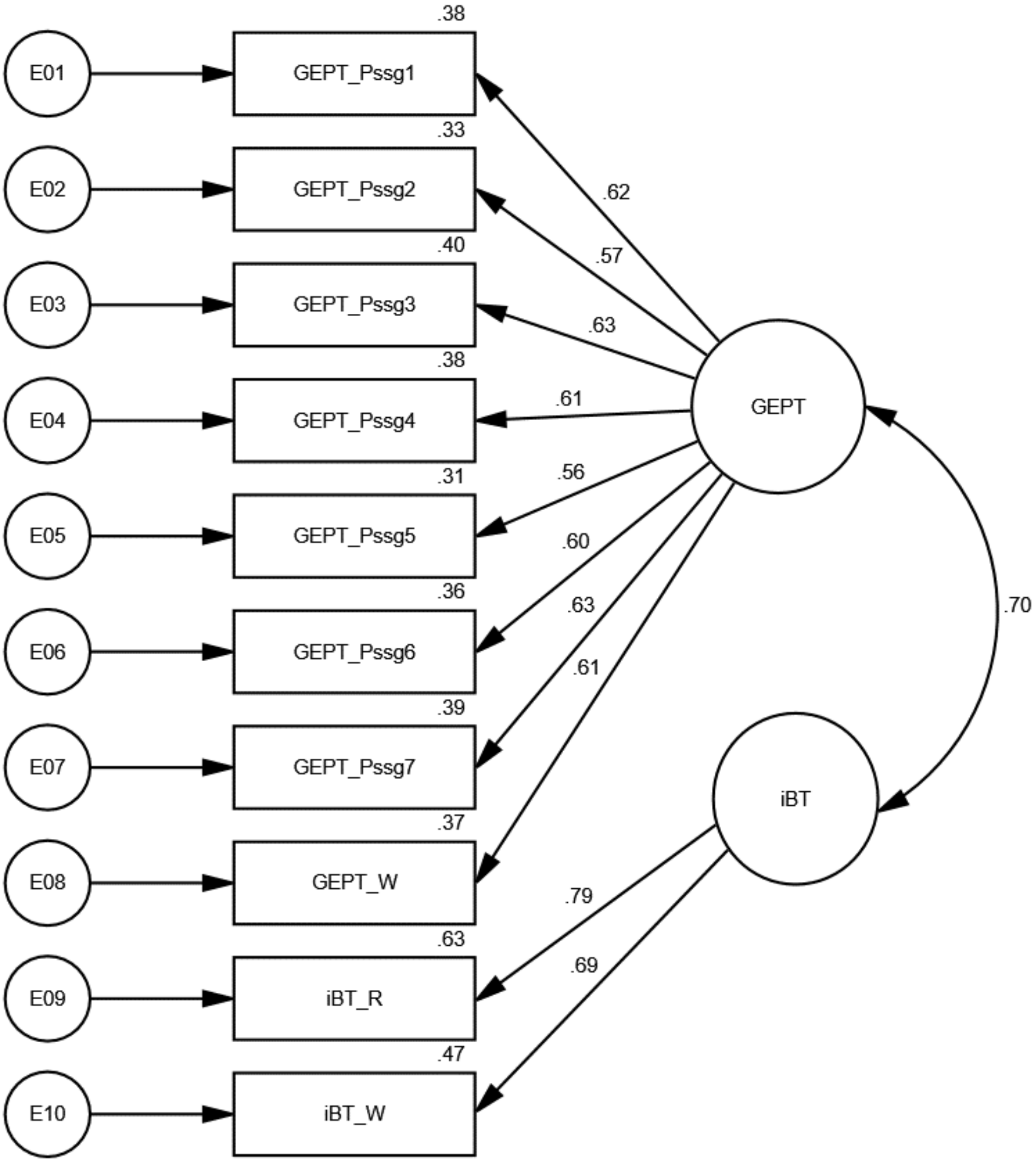

Figure 6. Path Diagram for Model 3 (2-Factor GEPT-A & iBT), with Parameter Estimates

Table 23. EFA Results for iBT Scores (Writing, Listening, Speaking & Reading)

Variables	Factor Loadings
iBT_W	.803
iBT_L	.823
iBT_S	.656
iBT_R	.664

Table 24. Goodness of Fit Summaries for GEPT-A-only and iBT-only CFAs

Statistic	Model 4	Model 5[a]	Model 6
χ^2	37.360	--	22.075
df	20	--	14
p	.011	--	.077
NFI	.903	--	.930
NNFI	.910	--	.959
CFI	.950	--	.972
RMSEA	.069	--	.056
RMSEA $CI_{.90}$	.033 - .103	--	.000 - .099

[a] Model 5 did not converge due to under-identification.

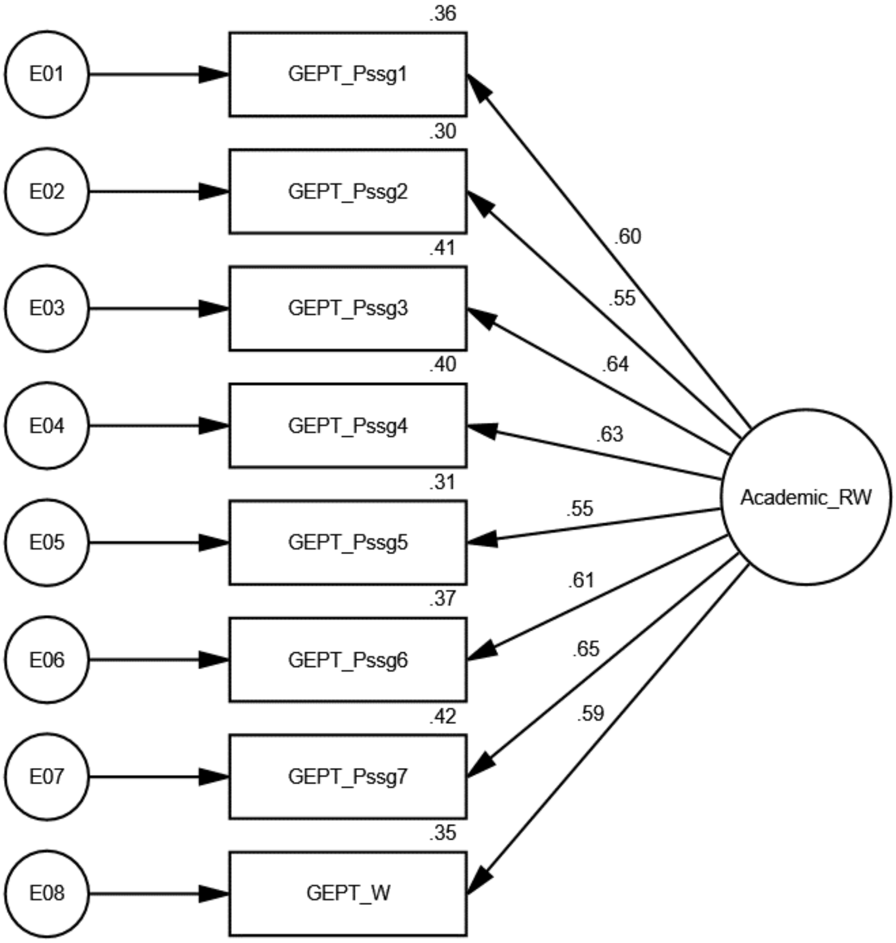

Figure 7. Path Diagram for Model 4 (1-Factor Reading & Writing—GEPT-A only), with Standardized Parameter Estimates

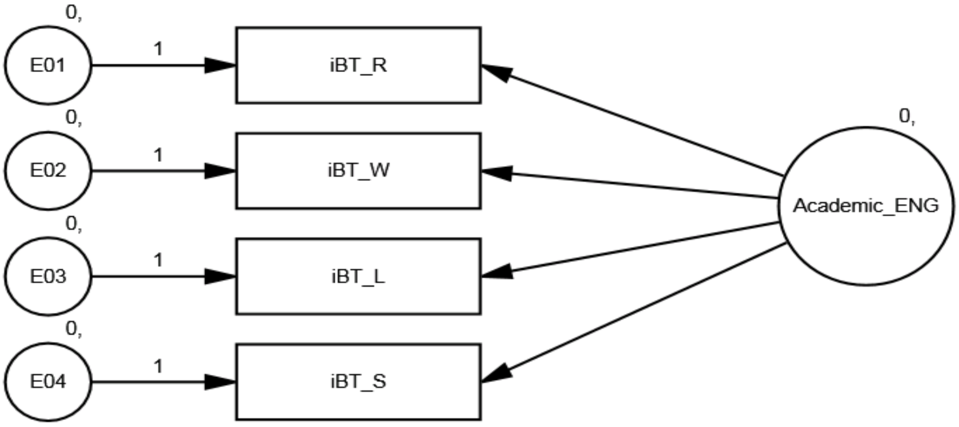

Figure 8. Path Diagram for Hypothesized Model 5 (1-Factor Academic English—iBT only), Which did not Converge due to Under-identification

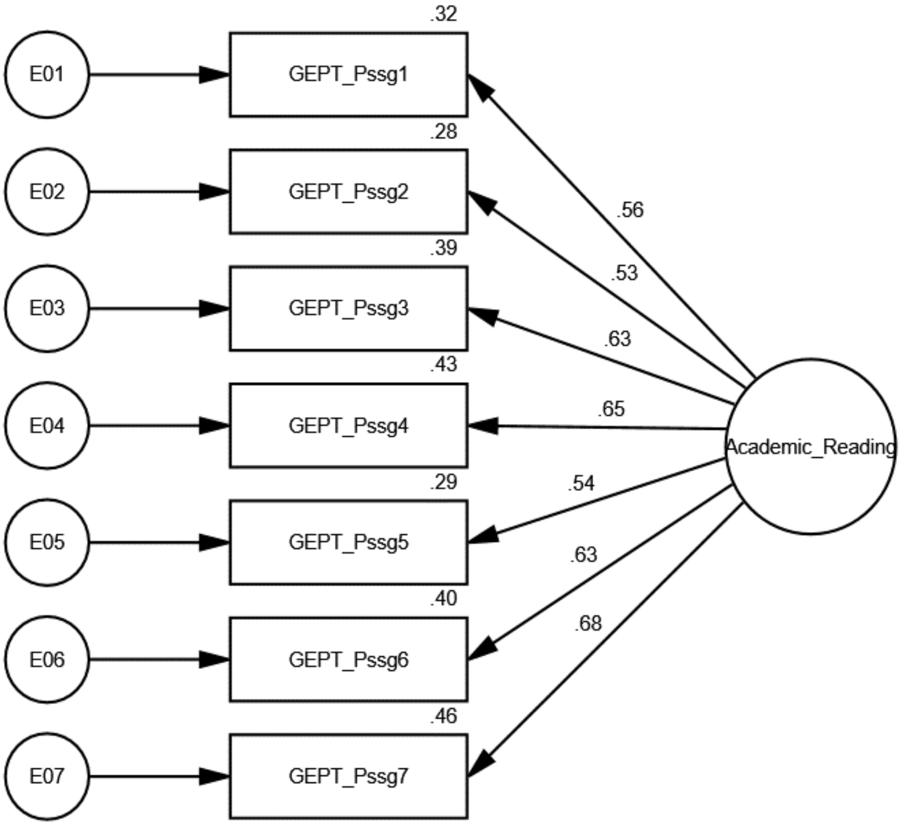

Figure 9. Path Diagram for Model 6 (1-Factor Reading—GEPT–A only), with Standardized Parameter Estimates

6. Discussion

In this section, we discuss the results described in the preceding section. We begin by addressing Research Question 1 with consideration of the results of the content analyses of the passages, followed by the task analysis of the reading items. We then take up Research Question 2 and the analysis of the test and survey results. We next consider Research Question 3 in light of the EFA and CFA results and the overall findings of the study. We conclude the section with a discussion of areas worthy of additional research.

6.1. Research Question 1: Content Analysis of Passages

Research Question 1 asked “What is the content of the reading and writing tests of the GEPT-A and the iBT – the test passages, items, tasks?” Aside from the obvious difference in length between the two tests (seven reading passages for the GEPT-A, four for careful reading and three for expeditious reading, vs. three for the iBT), the GEPT-A and iBT reading passages proved to be highly similar in most aspects. The most salient of the six significantly different features was the number of words per passage, with the GEPT-A passages averaging nearly

50 words more than the iBT reading passages (.6 standard deviations,[9] a medium effect size). The reading passages for the two tests also differed in certain aspects involving vocabulary. The GEPT-A had a higher level of lexical diversity in its passages (by 1.0 and 1.3 standard deviations) on two separate measures, clearly a large effect size. The proportion of words from the K1 list was also higher for the GEPT-A (by 1.1 standard deviations, also a large effect size). In syntax, the iBT had on average significantly more modifiers per noun phrase (1.1 standard deviations, a large effect size). Finally, in terms of syntactic similarity across paragraphs—an indicator of cohesion and/or of ease of processing—the iBT measured higher than the GEPT-A by 2.0 standard deviations (a large effect size).

We can therefore say that while they are similar in a host of other respects, the only significant differences between the reading passages on the two tests are that the GEPT is slightly longer, uses a markedly greater level of lexical variety, and uses more simple vocabulary—but not, oddly enough, a significantly lower level of more challenging vocabulary. In turn, the iBT features longer noun phrases, which presumably increase its syntactic complexity and the level of reading difficulty. The iBT also has much more consistent syntax at the sentence level than the GEPT-A, which should help increase cohesion while lowering reading difficulty. These variables are probably more important in determining the actual readability of a text for non-native speakers than are traditional readability measures (see, e.g., Carrell, 1987), even if the appropriate values for the alternative measures still await determination.

It is a limitation of the present study that the sample size for integrated writing input passages was so small. It seems likely that this was the main reason for no significant differences being identified for any of the input passage variables—even though, for example, the GEPT-A uses much longer input passages than the iBT. Further research with a greater number of passages from each test would be necessary to establish this conclusively, however.

6.1.1. Task analysis

In this section, which also relates to Research Question 1, we begin with a brief discussion of the topics used in the GEPT-A and iBT reading passages, comment on the construct coverage of the two reading tests (in terms of what aspects of reading they assess in the forms analyzed), and then discuss characteristics of the two tests in terms of the scope and task formats of the reading items.

Topics of the Reading Comprehension Passages. The GEPT-A form analyzed in this study included passages with a range of topical content, but had only one passage dealing with the sciences (and none taken from the life sciences). In contrast, the iBT had a much heavier emphasis on the sciences, with two science passages in two forms, and one in the remaining form. Interestingly, the only physical or Earth science topic covered by either the GEPT-A or iBT was geology (loosely defined, as one iBT passage dealt with icebergs). This may stem in part from the markedly different purposes of the two tests, in that many iBT test takers plan to study science or engineering in the United States, and the TOEFL therefore has had a long history of including reading passages from these subjects.

[9] Since pooled standard deviations were not used, this is not truly a Cohen's *d*, although the interpretation is similar.

Construct Coverage of the Reading Tests. The main difference between skimming a text and reading it for the gist is the degree of speededness of the task. Likewise, the difference between scanning and reading for specific details is primarily one of how rapidly the task is performed. The iBT did not include any skimming or scanning items; however, the GEPT-A skimming and scanning section included 20 questions (the same number as on the Careful Reading section), and was allocated a time limit of 20 minutes. Although the time limit for individual sections was not imposed on the U.S.-based test takers, their test instructions did say there was a 20-minute limit. We believe that this, along with the presence of the overall time limit, made it likely that test takers did in fact attempt these tasks in an expeditious fashion, and that they therefore probably skimmed and scanned, rather than reading carefully.
The GEPT-A did not include any questions targeting vocabulary knowledge or the ability to infer the meaning of unfamiliar vocabulary from context, while the iBT made heavy use of vocabulary questions (29.5% of all items). It should be pointed out that the iBT items that involved vocabulary could all be answered by a test taker unfamiliar with the words who was skilled at inferring the meaning of unfamiliar vocabulary from context. In fact, every one of the items could be answered using preexisting vocabulary knowledge instead. Given the task format used—multiple choice glosses of the word or phrase in question—examinees who knew the word could answer without even having to read the passage. Furthermore, in some cases, word analysis skills (e.g., application of knowledge of morphology) could be used to infer a definition without having to read the passage, as in Practice Test 1, where the word *immeasurably* was the focus of one item. By breaking the word down into *im-* + *measure* + *able* + *ly*, a strategic reader could identify the correct meaning without reading the passage. The existence of these alternatives to using the intended reading process is one of the main weaknesses of using this sort of task format to assess the ability to infer the meaning of unfamiliar vocabulary.

Furthermore, some vocabulary items on the iBT practice tests could not be answered except using prior vocabulary knowledge—that is, readers would not be able to determine the meaning of the target word from the context, and could not successfully answer the item unless they already knew the meaning of the word being tested. In summary, then, some of these iBT items would function as vocabulary-in-context items for test takers who did not already know the words, but would function as measures of vocabulary knowledge for anyone who already knew them.

Both tests include items that require students to read for specific details, and both tests make heavy use of paraphrasing rather than using identical language in an item and the passage—a practice that makes these items require more than the ability to simply identify relevant information in the passage, and something that would perhaps not be feasible with tests aimed at lower levels of language ability.[10] On the other hand, the careful reading sections of the GEPT-A make such extensive use of specific details items[11] (11 out of 20 careful reading questions) that there is little room left on the test for other aspects of the reading construct, such as inferencing or reading for the main idea.

Although both the GEPT-A and iBT include items assessing test takers' sensitivity to rhetorical organization, these items differ in important ways in terms of their scope. The one GEPT-A item assessing this portion of the reading construct had a very broad scope, meaning

[10] Indeed, it is possible that items with language paraphrased from the text might function similarly to inference items for low-proficiency test takers. At the level targeted by the GEPT-A, though, this is unlikely to be the case.
[11] Note, however, that LTTC considers several items to be assessing inferencing that we have classified as assessing specific details or paraphrasing/summarizing.

that answering it correctly required processing all or nearly all of the passage. On the other hand, most of the nine iBT items assessing this aspect of reading—one for every passage—required test takers to process all or nearly all of a single paragraph. Two items had narrow scope, with the key information needed to answer them spread across a few sentences. Only one rhetorical organization item had broad scope, requiring test takers to read more than one paragraph in order to answer correctly. None had very broad scope. Thus, the one item of this type on the GEPT was also the only one from either test to require attention to the rhetorical organization of the entire passage, rather than merely a portion of it.

Both tests required paraphrasing and summarizing of material read, but they differed in their emphasis. The GEPT-A requires both paraphrasing and summarizing, but with a greater emphasis on summarizing. In contrast, the iBT straddles the boundary between the two to some extent, and involves a much smaller degree of the information reduction that is required in summarizing. This stems at least in part from the difference in scope between the two tests for these items. In addition, the GEPT-A uses short-answer tasks to address this portion of the reading construct, while the iBT uses multiple choice. Furthermore, it is worth noting that *all* of the short-answer GEPT-A items—even those not targeting this portion of the reading construct—require at least some degree of paraphrasing because of the strict scoring rules regarding recycling of language taken directly from the passages (i.e., use of more than a key word or phrase is considered “plagiarism”).

Interestingly, the iBT appears to have abandoned main idea items in favor of major points. At the same time, however, the GEPT-A only included one main idea item and no major point items. Main idea questions have long been a standard part of testing reading, so it is surprising that the GEPT-A largely omits them. In discussing this finding, however, it is particularly important to keep in mind our determination that the GEPT-A skimming and scanning section did in fact require expeditious reading, not careful reading. Careful reading to identify the main idea of a paragraph would probably be equivalent to reading for major points, but the framework used here differentiates skimming (which inherently involves reading for the gist, and major ideas, of a text) as being qualitatively different from careful reading to identify the main idea (or major points) of a passage.

A final point in terms of the adequacy of the construct representation on these two reading tests involves inferencing and identifying author purpose. Arguably, the latter is an example of the former; in any case, the GEPT-A only includes one author purpose question on the form analyzed in this study, compared to eight inference items and nine author purpose items, or roughly three per test form, one per passage.

Scope. The reading comprehension questions on the GEPT-A and iBT differed substantially in terms of scope. The overwhelming majority of iBT items (76%) had narrow or very narrow scope—that is, the necessary information was contained within several sentences or just one sentence, respectively. In marked contrast, 75% of the GEPT-A reading questions on the form analyzed had moderate, broad, or very broad scope, requiring test takers to extract the necessary information from most or all of a paragraph, more than one paragraph, or more than half of the passage, respectively. This could be expected to make the GEPT-A questions more challenging overall, although verification of this is beyond the limits of the present study, given that iBT response data was not available.

Task formats. The two tests differed markedly in terms of the task formats they employed. The majority of reading items on the GEPT-A were selected response, but over a third were short answer items. The selected response items included a substantial portion that were not multiple choice—although about one third of all questions were multiple choice or fixed format multiple choice, the remaining 30% were matching. On the other hand, the iBT forms analyzed relied overwhelmingly on traditional multiple choice, with eight multiple-response multiple choice items and one categorization item spread across the nine passages. The limited use by the iBT of nontraditional or "enhanced" selected response task formats is better than none at all, but the more even distribution of task formats on the GEPT-A clearly sets it apart, and stands likely to reduce any impact from test method effects on the scores. Furthermore, the fairly heavy use of short-answer questions on the GEPT-A is more authentic (Bachman & Palmer, 1996); such items may engage communicative language ability more thoroughly, and could also do a better job of testing actual comprehension, as opposed to mere *recognition* of the correct answers in the options.

6.2. Research Question 2: Test Performance and Survey analyses

This section of the study relates to Research Question 2, which asked "What is the test performance of the study participants on the reading and writing tests of the GEPT-A at the total test score and at the item level, and on the iBT at the total test score level?" It answers this question by considering the overall scores on the GEPT-A and iBT, and the reliability and item performance of the GEPT-A. It then discusses the results of the test takers survey.

Descriptive statistics. Judging from the descriptive statistics for scores on the two tests, it appears that test takers got a higher proportion of questions correct on the iBT than on the GEPT-A—in fact, percentage correct scores were roughly 1.5 standard deviations higher on the iBT. Similarly, although they use very different rating scales and cannot be expected to be scaled the same, test takers seem to have done better on the iBT writing section than on the GEPT-A first writing task, in terms of percentage of points possible on the rating scale. Whether this is because the GEPT-A Writing Task 1 was rated more strictly than the overall iBT writing section (owing to either the rating scales themselves, rater training, or both), because the GEPT-A writing task was more difficult, or a combination of the two, cannot be determined from the data at hand.

The distribution of GEPT-A reading scores was close to normal, with minor negative skewedness and kurtosis. iBT reading scores, on the other hand, had high positive kurtosis, and were clearly more negatively skewed than GEPT-A reading scores, although not *severely* so. Similar but less extreme distributional patterns could be seen in the distributions for the other iBT sections, although they were only truly noteworthy in the case of listening. Given that the mean was only 1.2 standard deviations from the maximum possible score, this suggests that a ceiling effect was taking place in the iBT scores—particularly in the case of reading—at least with this population. If it is correct that there was a ceiling effect in the iBT scores, that fact could also be partially responsible for the low correlations between scores on the two tests, as a result of a restriction in range for the iBT scores. It should be noted, however, that such a ceiling effect might not be observed with a more typical sample, one more representative of the usual international iBT candidature, as opposed to the present sample, which had a higher overall level of language proficiency. This can be seen from the fact that the mean composite iBT score for participants in this study, 95.3, was equivalent to roughly the 73rd percentile among 2014 iBT test takers worldwide, and the mean reading score of 24.9 (82.9% of the possible scale points) was equivalent to roughly the 69th percentile (Educational Testing Service, 2015b).

Study participants tended to do about equally well on the two sections of the GEPT-A reading, the careful reading and skimming and scanning sections. Test takers' iBT scores were similar for reading, writing, and listening, but markedly lower for speaking. Reliability and the Standard Error of Measurement (SEM) for iBT reading typically average .85 and 3.35 out of 30 scale points (Educational Testing Service, 2011). This is roughly comparable to the results found for the GEPT-A reading in this study (.880, and 7.6 out of 120 scale points).

There were clearly noticeable differences in performance across the seven passages used in the GEPT-A. The first three passages were highly similar in format and the nature of their item formats, and the scores on the testlets (sets of items) associated with each passage were roughly comparable. Scores on the other testlets varied widely, with Passage 4 the most difficult, perhaps because of the nature of the task (summarizing and paraphrasing with short-answer questions). Most puzzling, however, was the marked difference between scores on Passages 5 and 6. These both required skimming, and were ostensibly quite similar, but for some reason scores differed on them by 15%. Any definitive statement as to the reason would probably require comparison with additional passages.

Item analysis. Most items had acceptable item analysis values. As the GEPT-A is a criterion-referenced test, item difficulty does not figure into judgments of item acceptability; however, it was reported for reference, and there were no terribly extreme cases, with only one item falling below .20 and only four above .80. As for discrimination, only six items were below the commonly used criterion of .30 on correlational discrimination indices for professionally developed items (see Carr, 2011). Only one was below .20; thus, the items were viewed as doing an adequate job of discrimination overall.

Survey analysis. Participants reported on average that they found the reading portion of the GEPT-A more difficult than the writing section, by about half of a standard deviation. For both portions of the test, "medium" difficulty was the most common description, with 10% more participants selecting that response for writing than did for reading. At the same time, three times as many participants found the reading "difficult" as did the writing. Twice as many rated the writing "easy" as gave that rating to the reading test.

The average rating for the relevance of test content to students' academic studies was equivalent to a rating of "neither agree nor disagree," for both GEPT-A reading and writing. The relevance of the reading content was judged higher than that of the writing content by about .40 standard deviations. The most common description chosen for the content relevance of both reading and writing tasks was "agree." Similarly, the average rating for the relevance of test tasks to participants' academic studies was equivalent to "disagree" for reading, and between that rating and "neither agree nor disagree" for writing. The most commonly selected rating for reading task relevance was "disagree," while the most common rating for writing task relevance was "agree."

In summary, the vast majority of participants found the GEPT-A reading tasks to be of high or medium difficulty, and similar numbers reported the writing tasks to be of low or medium difficulty. These results are puzzling, given that test takers actually tended to perform better on the reading than on the writing. Participants were generally neutral regarding the relevance of the reading test content and tasks to their academic studies, but most agreed that the content and tasks of the writing section were relevant.

6.3. Research Question 3: Comparability of the tests

This portion of the discussion addresses Research Question 3, which asked "What is the comparability of the reading and writing tests of the GEPT-A and iBT?" The pervasively significant correlations among the iBT reading and writing scores and GEPT-A reading testlets and writing scores (and the total GEPT-A reading scores as well) indicate that the two tests are in fact assessing related things.

Test takers' perceptions of the difficulty of the GEPT-A reading and writing tasks had a consistently negative relationship with their scores on the two tests. Most of these relationships were significant. This lends some additional—albeit weak—support to the idea that the two tests both assessed the same abilities.

The results of the EFA further indicated that the GEPT-A and iBT reading and writing sections measured substantially the same construct, since all observed variables (the seven passage-based GEPT-A testlets, GEPT-A writing score, and iBT reading and writing scores) loaded on the same common factor.

The hypothesis that the two tests measure the same constructs was also supported by the results of the CFA, although not with the same factor structure as suggested by the EFA. The CFA found the best fit for a two-factor (reading and writing) rather than for a single-factor model. The two-factor reading and writing model also fit better than one with separate factors for the GEPT and iBT. The two-factor reading and writing model did not fit the data as satisfactorily as could be hoped, though. This could have been because of problems with the model, particularly with the identification of the writing factor, which only loaded on two observed variables. Analyzing the results with a third writing variable might help this problem. It is also possible that the model was sufficiently identified, but that the size of the sample led in this case to empirical under-identification, a condition sometimes encountered in factor analytic studies in which a unique solution for all parameter estimates is not possible, despite the fact that all parameters are formally (i.e., algebraically) identified (Rindskopf, 1984). This latter possibility is supported by the fact that a single-factor model, which did not have the problem of a two-indicator factor, did not fit as well as the two-factor model.

Complicating the picture, however, is the fact that both the GEPT-A-only (Model 4) and GEPT-A reading-only (Model 6) models fit better than the best-fitting model that included both tests. This does not necessarily indicate that the two tests are assessing entirely different constructs. For model identification purposes, the iBT-only model had to include listening and speaking scores, whereas GEPT-A listening and speaking were not included. Considering the two models together, therefore, does not involve an apples-to-apples comparison. However, it does indicate that at least to some extent, the GEPT-A and iBT reading and writing sections are measuring somewhat different constructs. That is to say, both tests are clearly assessing reading and writing ability, but equally clearly, they appear to be assessing different aspects of the reading construct. Further support for this interpretation can be found in the strengths of the correlations between factors in the model with correlated reading and writing factors (Model 2a) and the model with correlated GEPT-A and iBT test factors (Model 3). In Model 2a, the reading and writing factors correlated at .85, whereas the GEPT-A and iBT factors in Model 3 correlated at .70. This indicates that at an overall level, the two tests differ even more than do the constructs of reading and writing. This interpretation is also supported by the findings of the task analysis regarding construct coverage, scope, and task formats. A well-fitting single model that includes both tests should be possible to construct, but would presumably require data at the same level (i.e., testlets from both tests, or individual items from both tests), and perhaps with a larger sample size than was obtained in this study.

As a final point, it is worth noting the surprising result that Model 5 (iBT scores only) would not converge, even though the four variables were so highly intercorrelated (r = .328 for the lowest value, and $r \geq .545$ for the other five correlations). This provides further indication that empirical under-identification was indeed a problem in the CFAs in this study. It may also relate to the potential ceiling effect identified in the iBT data for the present study; such a restriction of range might easily have impaired model fit for any of the models tested that included iBT data.

These results are parallel in some ways to those found in the comparability study mentioned in the review of literature: the Cambridge-TOEFL Comparability Study described above (Bachman et al., 1995). That study found that scores for individual sections of the FCE and TOEFL were sometimes more closely related to other sections of the same test than to sections from the other test that were intended to assess the same construct (i.e., sections that ostensibly assessed the same construct were often not as closely related to each other as they were to other sections of the same test, which were intended to assess other constructs). Similarly, the present study is not the first one in which clear and easily interpretable EFA results have been less clear, or even impossible to model, when replicated in CFA. For example, a higher-order factor model for which Bachman et al. (1995) had found a clear factor structure using EFA failed to converge at all when subsequently subjected to CFA by Kunnan (1995). It should be pointed out that these two previous studies employed larger samples than were used here, as well. Therefore, taken in context, the present results become somewhat less surprising.

6.4. Implications for Research Question 3 of Other Findings

The content analysis of the reading passages found that there are noticeable differences in the passages, but it is impossible to say from the results at hand how important the differences are in terms of effecting examinee performance, and the comparability of the two tests.[12] As for the topics used in the reading passages, based on the test form analyzed, the GEPT-A places much less emphasis on the reading of scientific or technical topics than the iBT. This is one area in which the two tests seem to *not* be comparable.

Given the contrasts between the two tests in terms of construct representation, it seems fair to say that the GEPT-A and iBT are not comparable in terms of the aspects of the larger reading construct that they assess. In particular, the GEPT-A does not give adequate coverage to aspects of careful reading besides reading for details and paraphrasing/summarizing, particularly inferencing and the ability to determine the meaning of unfamiliar vocabulary from context. At the same time, however, the iBT omits all coverage of skimming and scanning, and has too many items that function (or can function) as assessments of vocabulary knowledge rather than reading ability.

The tests are also not comparable in terms of the scope of their reading comprehension items. The GEPT-A seems to have a more even distribution in scope across its items than does the

[12] That would require a study comparing the performance of a single group of test takers on both the GEPT-A and iBT and analyzing the specific characteristics of the passages in use. Even so, identifying the effect of passage variables on test taker performance can be quite challenging (see, e.g., Carr, 2006).

iBT, with a much lower proportion of narrow-scope items than the iBT. This seems appropriate for a test that purports to assess English at a high level of proficiency, whereas a greater emphasis on items of narrow and very narrow scope would be appropriate on tests targeting lower proficiency levels. The task formats used on the two tests are also not comparable, most notably due to the extensive use of short answer questions on the GEPT-A.

In addition, the score distributions of the two tests were not equivalent in this study, suggesting that the GEPT-A may have been more difficult than the iBT. The reliability of the two tests was comparable, however. Similarly, the correlations between scores on the two tests were high enough to indicate that they probably assess related constructs, but were also low enough to make clear that there are marked differences as well—although, as noted above, ceiling effects in the iBT scores may have depressed the correlations between the two tests.

Furthermore, the results of the regression analyses indicated that while iBT reading and writing scores can be used to predict GEPT-A reading and Writing Task 1 scores, the relationship between the two sets of scores is tenuous due to small effect sizes. Based on the regression equations, however, as well as concordance tables published by ETS (Educational Testing Service, 2015a), C1 in the CEFR is equivalent to a 24 iBT reading or writing score, which equates to a GEPT-A reading score of 68, and a GEPT-A Writing Task 1 score of 3. Thus, these regression results must be interpreted with caution in view of small effect sizes and large standard errors of the estimates.

In conclusion, while the passage and task analyses revealed important differences between the two tests, the correlational analyses indicate that the GEPT-A and iBT are both assessing reading and writing, and the scores on the two are very closely related. It is probably most accurate to say that the two tests assess the same constructs but from somewhat different perspectives, and therefore with somewhat different construct definitions.

6.4.1. Areas for future research

One of the limitations of the present study is that it only involved the analysis of a single form of the GEPT-A. A replication of the task analysis from this study using additional GEPT forms would be desirable. This would show how representative this particular form was, and would provide a more reliable description of the test and the characteristics of its tasks. The use of additional iBT forms might be desirable as well, assuming they were balanced by equal numbers of GEPT-A forms.

This study found that reading scores were higher on the iBT than on the GEPT-A. The greater level of challenge for the GEPT-A could be due to some of the points identified in the task analysis, particularly the greater scope of most GEPT-A items and the fairly extensive use of limited production tasks, rather than a total reliance on selected response. Verification of this might be performed using a multi-trait multi-method study, a many-facet Rasch analysis, or preferably both methods used in conjunction.

A study of how various task and passage characteristics (including both Coh-Metrix output and vocabulary-related measures) might affect reading performance would be another interesting area of investigation. It was not possible in this study to explore this question, but a study with additional test forms and passage-based testlet difficulty placed on the same scale (e.g., estimated using IRT and anchor passages) might shed light on this topic. Certainly, empirical text measures that actually predicted testlet difficulty would be an invaluable resource for test development.

It would be desirable to attempt a CFA of the GEPT-A with a larger sample size and separate scores for each of the subscales on the analytic rating scale, and perhaps with listening and speaking scores as well. A clearer understanding of the factor structure of the GEPT-A would be a useful component in the overall validity argument for the test.

7. Conclusion

This study aimed to investigate the comparability of the GEPT-A and iBT using data from test takers in both Taiwan and the United States, a form of the GEPT-A reading and writing sections, and iBT reading and writing test forms published commercially by ETS. Three research questions were posed regarding the content of the GEPT-A and iBT reading and writing tests, performance on the two tests, and the comparability of the two tests.

We concluded that the passages on the two tests are comparable in many ways, but reading passages differ in several key regards. The task analysis revealed that the construct coverage, item scope, and task formats of the two tests are clearly distinct. Analysis of participant responses indicated that the GEPT-A has good reliability, and that reading comprehension items tend to function quite well. It also appears that the two tests assess the same constructs, but emphasize different *aspects* of the reading construct, making results on the two tests not entirely comparable.

References

Bachman, L. F., Davidson, F., Ryan, K., & Choi, I-C. (1995). *An investigation of the comparability of the two tests of English as a foreign language: The Cambridge-TOEFL comparability study.* Cambridge, U.K.: Cambridge University Press.

Bachman, L. F., Davidson, F., & Milanovic, M. (1996). The use of test methods in the content analysis and design of EFL proficiency tests. *Language Testing, 13*, 125-50.

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

Bachman, L. F., & Palmer, A. S. (2010). *Language assessment in practice.* Oxford: Oxford University Press.

Bridgeman, B. & Cooper, R. (1998). *Comparability of Scores on Word-Processed and Handwritten Essays on the Graduate Management Admissions Test.* Research report 143: Educational Testing Service, Princeton, NJ.

Carr, N. T. (2006). The factor structure of test task characteristics and examinee performance. *Language Testing, 23*, 269-289.

Carr, N. T. (2011). *Designing and analyzing language tests*. Oxford: Oxford University Press.

Carrell, P. L. (1987). Readability in ESL. *Reading in a Foreign Language, 4,* 21-40.

Choi, I-C., Kim, K. S., & Boo, J. (2003). Comparability of a paper-based language test and a computer-based language test. *Language Testing, 20,* 295-320.

Cobb, T. (n.d.). *Web vocabprofile.* Retrieved from http://www.lextutor.ca/vp/comp/

Coh-Metrix. (n.d.). Coh-Metrix web tool. Retrieved from http://www.cohmetrix.com/

Comrey, A. L., & Lee, H. B. (1992). *A first course in factor analysis.* Hillsdale, NJ: Lawrence Erlbaum Associates.

Coxhead, A. (2000). A new academic word list. *TESOL Quarterly, 34*, 213-238.

Crossley, S. A., Dufty, D. F., McCarthy, P. M., & McNamara, D. S. (2007). Toward a new readability: A mixed model approach. In *Proceedings of the 29th annual conference of the Cognitive Science Society* (pp. 197-202).

Educational Testing Service. (2011). *Reliability and comparability of TOEFL iBT scores* (TOEFL iBT Research Insight Series 1, Vol. 3). Retrieved from http://www.ets.org/s/toefl/pdf/toefl_ibt_research_s1v3.pdf

Educational Testing Service. (2012). *The official guide to the TOEFL test* (4th Ed.). McGraw-Hill.

Educational Testing Service. (2015a). *Compare TOEFL scores.* Retrieved from:

http://www.ets.org/toefl/institutions/scores/compare/

Educational Testing Service. (2015b). *Test and score data summary for TOEFL iBT Tests: January 2014 – December 2014 test data.* Retrieved from http://www.ets.org/s/toefl/pdf/94227_unlweb.pdf

Geranpayeh, A. (1994). Are Score Comparisons Across Language Proficiency Test Batteries Justified? An IELTS-TOEFL Comparability Study. *Edinburgh Working Papers in Applied Linguistics, 5,* 50-65.

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

Halliday, M.A.K, & Hasan, R. (1976). *Cohesion in English*. London: Longman.

Hawkey, R. (2009). *Examining FCE and CAE*. Cambridge, UK: Cambridge University Press.

Hu, L., & Bentler, P. M. (1995). Evaluating model fit. In Rick H. Hoyle (ed.), *Structural equation modeling: Concepts, issues, and applications*. Sage Publications: Thousand Oaks, CA.

Kunnan, A. J. (1995). *Test taker characteristics and test performance: A structural modeling study.* Cambridge, UK: Cambridge University Press.

Kunnan, A. J. (1998). An introduction to structural equation modeling for language assessment research. *Language Testing, 15*, 295-332.

Kunnan, A. J. (2012). *Portability and the CEFR.* Talk given at the Language Testing Research Colloquium, Princeton, NJ.

O'Loughlin, K. (1997). *The comparability of direct and semi-direct speaking tests: a case study.* Unpublished Ph.D. thesis. University of Melbourne.

Rindskopf, D. (1984). Structural equation models: empirical identification, Heywood cases, and related problems. *Sociological Methods & Research, 13,* 109-119.

Sawaki, Y. (2001). Comparability of conventional and computerized tests of reading in a second language. *Language Learning & Technology, 5*, 38-59.

Stansfield, C. & Kenyon, D. (1992). Research on the comparability of the oral proficiency interview and the simulated oral proficiency interview. *System, 20*, 347-364.

Weir, C. & Wu, R. (2006). Establishing test form and individual task comparability: a case study of a semi-direct speaking test. *Language Testing, 23*, 167-197.

West, M. (1953). *A general service list of English words.* London: Longman, Green and Co.

Appendices

Appendix A

Table A1. Coh-Metrix Analyses of GEPT-A Reading Passages

Index	GEPT average[a]	GEPT R1	GEPT R2	GEPT R3	GEPT R4	GEPT R4b	GEPT R5	GEPT R6	GEPT R7	GEPT R7a	GEPT R7b	GEPT R7c
DESPC (Paragraph count, number of paragraphs)	6.7	6	6	10	6	1	8	8	16	3	3	4
DESSC (Sentence count, number of sentences)	36.8	37	31	41	37	8	48	51	49	13	15	15
DESWC (Word count, number of words)	685.0	731	620	725	756	162	863	844	910	285	299	310
DESPL (Paragraph length, number of sentences, mean)	5.5	6.2	5.2	4.1	6.2	8	6.0	6.4	3.1	4.3	5.0	3.8
DESSL (Sentence length, number of words, mean)	19.2	19.8	20.0	17.7	20.4	20.25	18.0	16.5	18.6	21.9	19.9	20.7
DESWLsy (Word length, number of syllables, mean)	1.7	1.7	1.7	1.7	1.7	1.7	1.7	1.6	1.7	1.8	1.8	1.6
DESWLlt (Word length, number of letters, mean)	5.2	5.0	5.4	5.2	5.3	5.2	5.2	5.0	5.4	5.5	5.6	5.2

Index	GEPT average[a]	GEPT R1	GEPT R2	GEPT R3	GEPT R4	GEPT R4b	GEPT R5	GEPT R6	GEPT R7	GEPT R7a	GEPT R7b	GEPT R7c
PCNARp (Text Easability PC Narrativity, percentile)	24.2	29.8	37.8	16.1	15.2	20.9	18.4	47.2	7.8	8.7	7.9	5.4
PCSYNp (Text Easability PC Syntactic simplicity, percentile)	48.6	40.1	27.8	62.6	40.5	40.5	62.6	68.1	56.8	33.7	62.6	28.1
PCCNCp (Text Easability PC Word concreteness, percentile)	75.1	87.7	61.0	55.6	73.6	37.5	84.1	66.3	97.3	91.2	97.1	99.8
PCREFp (Text Easability PC Referential cohesion, percentile)	44.8	31.9	76.1	58.7	28.1	2.39	36.7	46.8	15.4	42.5	3.9	34.5
PCDCp (Text Easability PC Deep cohesion, percentile)	68.7	88.9	74.9	65.2	81.9	54.0	57.9	86.9	25.1	33.7	25.5	16.1
PCCONNp (Text Easability PC Connectivity, percentile)	3.3	4.7	0.5	1.1	2.4	3.4	12.9	0.5	1.0	0.3	4.6	0.3
LDTTRc (Lexical diversity, type-token ratio, content word lemmas)	0.7	0.759	0.601	0.601	0.684	0.897	0.691	0.68	0.716	0.789	0.836	0.801
LDTTRa (Lexical diversity, type-token ratio, all words)	0.5	0.503	0.429	0.443	0.491	0.691	0.488	0.458	0.5	0.614	0.638	0.599

Index	GEPT average[a]	GEPT R1	GEPT R2	GEPT R3	GEPT R4	GEPT R4b	GEPT R5	GEPT R6	GEPT R7	GEPT R7a	GEPT R7b	GEPT R7c
LDMTLDa (Lexical diversity, MTLD, all words)	106.0	107.7	90.5	93.7	131.6	162	110.8	100.4	127.0	104.4	139.2	125.0
LDVOCDa (Lexical diversity, VOCD, all words)	106.9	102.5	104.9	97.0	124.4	117.7	104.8	105.1	125.5	108.8	120.2	101.4
CNCAll (All connectives incidence)	89.4	104.0	85.5	80.0	95.2	74.1	73.0	97.2	76.9	91.2	70.2	74.2
SYNLE (Left embeddedness, words before main verb, mean)	5.2	5.1	4.2	5.4	6.0	5.1	6.1	4.7	4.6	4.9	4.7	5.9
SYNNP (Number of modifiers per noun phrase, mean)	0.9	0.9	0.8	0.9	0.9	0.7	1.0	0.7	1.1	1.0	1.0	1.3
SYNSTRUTa (Sentence syntax similarity, adjacent sentences, mean)	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1
SYNSTRUTt (Sentence syntax similarity, all combinations, across paragraphs, mean)	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1
DRNP (Noun phrase density, incidence)	383.8	400.8	374.2	362.8	371.7	370.4	385.9	362.6	397.8	421.1	384.6	374.2

Index	GEPT average[a]	GEPT R1	GEPT R2	GEPT R3	GEPT R4	GEPT R4b	GEPT R5	GEPT R6	GEPT R7	GEPT R7a	GEPT R7b	GEPT R7c
DRVP (Verb phrase density, incidence)	188.9	158.7	193.5	233.1	199.7	265.4	177.3	220.4	151.6	147.4	170.6	145.2
DRAP (Adverbial phrase density, incidence)	31.7	34.2	35.5	27.6	39.7	49.4	27.8	49.8	14.3	10.5	20.1	12.9
DRPP (Preposition phrase density, incidence)	129.2	156.0	112.9	126.9	129.6	117.3	130.9	117.3	123.1	129.8	143.8	100.0
DRPVAL (Agentless passive voice density, incidence)	11.1	9.6	8.1	16.5	7.9	12.3	8.1	13.0	15.4	14.0	16.7	16.1
DRNEG (Negation density, incidence)	5.0	1.4	9.7	2.8	2.6	6.2	3.5	11.8	1.1	3.5	0.0	0.0
DRGERUND (Gerund density, incidence)	17.1	16.4	14.5	23.4	26.5	24.7	18.5	15.4	12.1	7.0	20.1	9.7
DRINF (Infinitive density, incidence)	16.9	10.9	16.1	22.1	18.5	30.9	18.5	26.1	9.9	7.0	16.7	6.5
WRDPRO (Pronoun incidence)	32.4	53.4	54.8	17.9	23.8	24.7	27.8	46.2	15.4	10.5	20.1	16.1
WRDAOAc (Age of acquisition for content words, mean)	368.1	357.8	363.2	357.9	385.2	373.6	386.7	367.0	338.4	357.1	332.0	327.6
WRDFAMc (Familiarity for content words, mean)	560.7	557.2	575.2	562.9	560.6	567.4	545.9	565.9	562.3	561.0	563.5	563.6

Index	GEPT average[a]	GEPT R1	GEPT R2	GEPT R3	GEPT R4	GEPT R4b	GEPT R5	GEPT R6	GEPT R7	GEPT R7a	GEPT R7b	GEPT R7c
WRDCNCc (Concreteness for content words, mean)	397.8	402.0	373.1	396.6	386.6	355.1	414.0	387.6	447.0	415.4	451.3	466.0
WRDIMGc (Imagability for content words, mean)	425.2	431.1	402.9	414.7	413.6	387.2	439.1	422.5	473.5	443.2	478.6	492.6
WRDPOLc (Polysemy for content words, mean)	3.6	3.3	3.9	4.0	3.9	4.4	3.6	3.6	3.0	3.2	3.1	2.7
RDFRE (Flesch Reading Ease)	43.5	45.0	40.3	47.0	41.2	42.6	47.1	51.7	41.5	32.3	37.5	48.6
RDFKGL (Flesch-Kincaid Grade Level)	12.0	11.9	12.6	11.1	12.6	12.3	11.2	10.2	12.1	14.2	13.0	11.6
RDL2 (Coh-Metrix L2 Readability)	12.6	11.2	16.5	14.0	14.8	11.3	10.2	14.2	8.8	9.1	9.3	7.4

[a] Average for the seven passages, weighted by number of words.

Table A2. Coh-Metrix Analyses of iBT Reading Passages

Index	iBT average[a]	iBT R1	iBT R2	iBT R3	iBT R4	iBT R5	iBT R6	iBT R7	iBT R8	iBT R9
DESPC (Paragraph count, number of paragraphs)	6.2	7	7	6	4	6	7	6	7	6
DESSC (Sentence count, number of sentences)	34.3	39	37	30	38	27	29	38	38	32
DESWC (Word count, number of words)	687.4	681	679	621	726	704	667	710	668	718
DESPL (Paragraph length, number of sentences, mean)	5.7	5.6	5.3	5.0	9.5	4.5	4.1	6.3	5.4	5.3
DESSL (Sentence length, number of words, mean)	20.4	17.5	18.4	20.7	19.1	26.1	23.0	18.7	17.6	22.4

Index	iBT average[a]	iBT R1	iBT R2	iBT R3	iBT R4	iBT R5	iBT R6	iBT R7	iBT R8	iBT R9
DESWLsy (Word length, number of syllables, mean)	1.7	1.8	1.7	1.6	1.5	1.7	1.6	1.7	1.8	1.7
DESWLlt (Word length, number of letters, mean)	5.1	5.4	5.3	4.9	4.8	5.1	4.9	5.2	5.1	5.1
PCNARp (Text Easability PC Narrativity, percentile)	18.2	20.6	9.9	9.7	25.1	44.0	16.9	9.5	14.2	12.3
PCSYNp (Text Easability PC Syntactic simplicity, percentile)	44.5	62.2	67.4	44.8	50.4	9.5	23.0	49.6	72.9	22.4
PCCNCp (Text Easability PC Word concreteness, percentile)	67.0	40.9	54.8	90.5	48.8	88.3	98.3	79.4	13.1	89.3

Index	iBT average[a]	iBT R1	iBT R2	iBT R3	iBT R4	iBT R5	iBT R6	iBT R7	iBT R8	iBT R9
PCREFp (Text Easability PC Referential cohesion, percentile)	40.2	25.1	45.2	26.1	24.2	35.2	93.9	36.7	31.6	44.4
PCDCp (Text Easability PC Deep cohesion, percentile)	54.4	28.8	66.6	57.1	44.0	87.3	21.5	83.7	65.9	33.7
PCCONNp (Text Easability PC Connectivity, percentile)	2.1	0.0	2.9	0.4	1.7	0.8	0.2	1.0	3.5	8.2
LDTTRc (Lexical diversity, type-token ratio, content word lemmas)	0.7	0.7	0.6	0.7	0.6	0.8	0.6	0.7	0.6	0.7
LDTTRa (Lexical diversity, type-token ratio, all words)	0.5	0.5	0.4	0.5	0.4	0.5	0.4	0.5	0.4	0.5

Index	iBT average[a]	iBT R1	iBT R2	iBT R3	iBT R4	iBT R5	iBT R6	iBT R7	iBT R8	iBT R9
LDMTLDa (Lexical diversity, MTLD, all words)	84.5	79.7	76.0	105.2	70.0	132.8	65.7	85.1	68.1	79.1
LDVOCDa (Lexical diversity, VOCD, all words)	86.7	77.0	98.1	95.8	83.4	106.8	69.8	88.1	85.5	76.4
CNCAll (All connectives incidence)	86.5	94.0	91.3	90.2	86.8	102.3	73.5	94.4	79.3	66.9
SYNLE (Left embeddedness, words before main verb, mean)	4.8	4.7	5.9	4.7	5.0	3.4	5.3	3.9	5.1	5.3
SYNNP (Number of modifiers per noun phrase, mean)	1.1	0.9	1.0	1.0	1.2	1.0	1.3	1.0	0.9	1.2

Index	iBT average[a]	iBT R1	iBT R2	iBT R3	iBT R4	iBT R5	iBT R6	iBT R7	iBT R8	iBT R9
SYNSTRUTa (Sentence syntax similarity, adjacent sentences, mean)	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1
SYNSTRUTt (Sentence syntax similarity, all combinations, across paragraphs, mean)	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1
DRNP (Noun phrase density, incidence)	376.1	397.9	381.4	360.7	355.4	420.5	352.3	390.1	363.8	360.7
DRVP (Verb phrase density, incidence)	173.8	182.1	185.6	182.0	176.3	127.8	167.9	188.7	179.6	175.4
DRAP (Adverbial phrase density, incidence)	27.5	20.6	19.1	40.3	35.8	28.4	18.0	31.0	24.0	30.6

Index	iBT average[a]	iBT R1	iBT R2	iBT R3	iBT R4	iBT R5	iBT R6	iBT R7	iBT R8	iBT R9
DRPP (Preposition phrase density, incidence)	133.3	129.2	138.4	132.0	112.9	142.0	134.9	131.0	142.2	137.9
DRPVAL (Agentless passive voice density, incidence)	12.6	5.9	8.8	20.9	4.1	7.1	19.5	16.9	18.0	13.9
DRNEG (Negation density, incidence)	4.0	5.9	2.9	0.0	4.1	2.8	7.5	2.8	9.0	1.4
DRGERUND (Gerund density, incidence)	16.4	23.5	20.6	16.1	13.8	9.9	15.0	14.1	4.5	29.2
DRINF (Infinitive density, incidence)	11.5	20.6	7.4	11.3	9.6	2.8	9.0	15.5	12.0	15.3
WRDPRO (Pronoun incidence)	31.4	36.7	20.6	16.1	37.2	85.2	15.0	29.6	19.5	19.5

Index	iBT average[a]	iBT R1	iBT R2	iBT R3	iBT R4	iBT R5	iBT R6	iBT R7	iBT R8	iBT R9
WRDAOAc (Age of acquisition for content words, mean)	365.9	420.9	408.0	334.9	347.6	366.5	315.5	378.9	403.4	317.5
WRDFAMc (Familiarity for content words, mean)	560.3	563.5	552.4	563.3	568.1	557.4	552.8	562.8	554.965	566.1
WRDCNCc (Concreteness for content words, mean)	399.1	368.9	375.5	424.8	386.7	400.0	452.6	409.2	364.2	412.7
WRDIMGc (Imagability for content words, mean)	427.8	410.7	417.9	449.6	412.0	423.8	481.7	427.1	393.1	437.3
WRDPOLc (Polysemy for content words, mean)	3.7	3.4	3.5	4.0	3.6	3.3	4.3	4.4	3.8	3.5
RDFRE (Flesch Reading Ease)	44.4	34.5	42.5	52.3	56.5	39.1	51.9	45.2	39.3	38.9

Index	iBT average[a]	iBT R1	iBT R2	iBT R3	iBT R4	iBT R5	iBT R6	iBT R7	iBT R8	iBT R9
RDFKGL (Flesch-Kincaid Grade Level)	12.1	12.8	11.9	11.1	10.1	14.3	11.7	11.6	12.1	13.4
RDL2 (Coh-Metrix L2 Readability)	9.8	10.9	10.4	7.9	12.0	7.4	10.8	9.5	12.7	6.9

[a] Average for the nine passages, weighted by number of words.

Table A3. Coh-Metrix Analysis of Input Passages for GEPT-A Independent Writing Task (W1)

Index	GEPT Average[a]	GEPT WR1	GEPT WR2
DESPC (Paragraph count, number of paragraphs)	5	5	5
DESSC (Sentence count, number of sentences)	20	20	20
DESWC (Word count, number of words)	413.5	410	417
DESPL (Paragraph length, number of sentences, mean)	4	4	4
DESSL (Sentence length, number of words, mean)	20.7	20.5	20.9
DESWLsy (Word length, number of syllables, mean)	1.8	1.8	1.7
DESWLlt (Word length, number of letters, mean)	5.4	5.6	5.2
PCNARp (Text Easability PC Narrativity, percentile)	20.5	19.8	21.2
PCSYNp (Text Easability PC Syntactic simplicity, percentile)	38.7	30.5	46.8
PCCNCp (Text Easability PC Word concreteness, percentile)	53.6	52.8	54.4

Index	GEPT Average[a]	GEPT WR1	GEPT WR2
PCREFp (Text Easability PC Referential cohesion, percentile)	56.5	73.2	40.1
PCDCp (Text Easability PC Deep cohesion, percentile)	75.2	55.2	94.8
PCCONNp (Text Easability PC Connectivity, percentile)	0.4	0.8	0.0
LDTTRc (Lexical diversity, type-token ratio, content word lemmas)	0.7	0.6	0.7
LDTTRa (Lexical diversity, type-token ratio, all words)	0.5	0.5	0.5
LDMTLDa (Lexical diversity, MTLD, all words)	91.7	80.6	102.6
LDVOCDa (Lexical diversity, VOCD, all words)	93.6	75.0	111.9
CNCAll (All connectives incidence)	100.4	95.1	105.5
SYNLE (Left embeddedness, words before main verb, mean)	5.4	5.5	5.3
SYNNP (Number of modifiers per noun phrase, mean)	1.0	1.0	1.0
SYNSTRUTa (Sentence syntax similarity, adjacent sentences, mean)	0.1	0.1	0.1

Index	GEPT Average[a]	GEPT WR1	GEPT WR2
SYNSTRUTt (Sentence syntax similarity, all combinations, across paragraphs, mean)	0.1	0.1	0.1
DRNP (Noun phrase density, incidence)	354.3	356.1	352.5
DRVP (Verb phrase density, incidence)	197.1	200.0	194.2
DRAP (Adverbial phrase density, incidence)	32.6	41.5	24.0
DRPP (Preposition phrase density, incidence)	114.9	109.8	119.9
DRPVAL (Agentless passive voice density, incidence)	4.8	2.4	7.2
DRNEG (Negation density, incidence)	10.9	12.2	9.6
DRGERUND (Gerund density, incidence)	24.2	22.0	26.4
DRINF (Infinitive density, incidence)	24.2	26.8	21.6
WRDPRO (Pronoun incidence)	27.8	26.8	28.8
WRDAOAc (Age of acquisition for content words, mean)	415.5	410.1	420.8
WRDFAMc (Familiarity for content words, mean)	559.9	560.6	559.2
WRDCNCc (Concreteness for content words, mean)	375.6	379.1	372.1

Index	GEPT Average[a]	GEPT WR1	GEPT WR2
WRDIMGc (Imagability for content words, mean)	399.6	403.8	395.5
WRDPOLc (Polysemy for content words, mean)	3.9	3.9	4.0
RDFRE (Flesch Reading Ease)	36.7	31.5	41.9
RDFKGL (Flesch-Kincaid Grade Level)	13.3	14.0	12.6
RDL2 (Coh-Metrix L2 Readability)	11.7	11.9	11.6

[a] Average for the two passages, weighted by number of words.

Table A4. Coh-Metrix Analysis of Input Passages (Reading and Listening) for iBT Integrated Writing Tasks

Index	iBT Reading Average[a]	iBT WR1	iBT WR2	iBT WR3	iBT Listening Average[b]	iBT WL1	iBT WL2	iBT WL3
DESPC (Paragraph count, number of paragraphs)	4.0	3	4	5	3.3	3	3	4
DESSC (Sentence count, number of sentences)	16.8	13	20	17	16.9	13	20	17
DESWC (Word count, number of words)	284.9	267	288	298	297.0	268	315	304
DESPL (Paragraph length, number of sentences, mean)	4.2	4.333	5	3.4	5.1	4.333	6.667	4.25
DESSL (Sentence length, number of words, mean)	17.4	20.538	14.4	17.529	18.0	20.615	15.75	17.882

Index	iBT Reading Average[a]	iBT WR1	iBT WR2	iBT WR3	iBT Listening Average[b]	iBT WL1	iBT WL2	iBT WL3
DESWLsy (Word length, number of syllables, mean)	1.6	1.588	1.865	1.473	1.5	1.56	1.575	1.477
DESWLlt (Word length, number of letters, mean)	4.9	4.659	5.444	4.654	4.6	4.709	4.689	4.467
PCNARp (Text Easability PC Narrativity, percentile)	31.7	34.46	21.48	38.97	42.6	42.07	60.64	24.51
PCSYNp (Text Easability PC Syntactic simplicity, percentile)	55.7	26.76	74.54	63.31	49.1	44.83	56.75	44.83
PCCNCp (Text Easability PC Word concreteness, percentile)	59.3	77.94	50.8	50.8	42.2	30.15	6.43	89.97

Index	iBT Reading Average[a]	iBT WR1	iBT WR2	iBT WR3	iBT Listening Average[b]	iBT WL1	iBT WL2	iBT WL3
PCREFp (Text Easability PC Referential cohesion, percentile)	59.6	53.19	48.8	75.8	36.7	6.18	29.46	71.23
PCDCp (Text Easability PC Deep cohesion, percentile)	90.5	89.8	96.86	84.85	81.7	81.86	95.64	67
PCCONNp (Text Easability PC Connectivity, percentile)	2.4	7.21	0.06	0.33	1.1	0.08	2.44	0.52
LDTTRc (Lexical diversity, type-token ratio, content word lemmas)	0.7	0.709	0.661	0.641	0.7	0.844	0.66	0.598
LDTTRa (Lexical diversity, type-token ratio, all words)	0.5	0.521	0.493	0.482	0.5	0.617	0.525	0.434

Index	iBT Reading Average[a]	iBT WR1	iBT WR2	iBT WR3	iBT Listening Average[b]	iBT WL1	iBT WL2	iBT WL3
LDMTLDa (Lexical diversity, MTLD, all words)	74.0	89	69.115	65.219	71.2	96.937	76.012	43.659
LDVOCDa (Lexical diversity, VOCD, all words)	77.5	80.867	80.088	72.08	84.1	114.463	92.04	49.14
CNCAll (All connectives incidence)	99.6	93.633	125	80.537	97.0	104.478	92.063	95.395
SYNLE (Left embeddedness, words before main verb, mean)	3.8	4.923	3.6	3.118	3.7	2.692	4.35	3.941
SYNNP (Number of modifiers per noun phrase, mean)	0.7	0.621	0.724	0.861	1.0	1.016	0.768	1.141

Index	iBT Reading Average[a]	iBT WR1	iBT WR2	iBT WR3	iBT Listening Average[b]	iBT WL1	iBT WL2	iBT WL3
SYNSTRUTa (Sentence syntax similarity, adjacent sentences, mean)	0.1	0.073	0.136	0.076	0.1	0.045	0.063	0.049
SYNSTRUTt (Sentence syntax similarity, all combinations, across paragraphs, mean)	0.1	0.062	0.113	0.087	0.1	0.039	0.051	0.063
DRNP (Noun phrase density, incidence)	410.3	430.712	416.667	385.906	338.2	291.045	355.556	361.842
DRVP (Verb phrase density, incidence)	194.6	198.502	159.722	224.832	206.3	220.149	241.27	157.895
DRAP (Adverbial phrase density, incidence)	35.2	18.727	41.667	43.624	42.8	67.164	31.746	32.895

Index	iBT Reading Average[a]	iBT WR1	iBT WR2	iBT WR3	iBT Listening Average[b]	iBT WL1	iBT WL2	iBT WL3
DRPP (Preposition phrase density, incidence)	154.7	157.303	170.139	137.584	107.1	78.358	104.762	134.868
DRPVAL (Agentless passive voice density, incidence)	14.1	3.745	10.417	26.846	12.4	11.194	3.175	23.026
DRNEG (Negation density, incidence)	9.4	0	6.944	20.134	15.8	11.194	28.571	6.579
DRGERUND (Gerund density, incidence)	16.4	22.472	20.833	6.711	15.8	14.925	25.397	6.579
DRINF (Infinitive density, incidence)	14.1	22.472	10.417	10.067	16.9	18.657	25.397	6.579
WRDPRO (Pronoun incidence)	41.0	48.689	38.194	36.913	25.9	26.119	31.746	19.737

Index	iBT Reading Average[a]	iBT WR1	iBT WR2	iBT WR3	iBT Listening Average[b]	iBT WL1	iBT WL2	iBT WL3
WRDAOAc (Age of acquisition for content words, mean)	365.8	341.512	469.795	287	358.0	354.636	437.325	278.795
WRDFAMc (Familiarity for content words, mean)	565.6	561.626	569.875	564.922	570.1	570.385	579.4	560.129
WRDCNCc (Concreteness for content words, mean)	401.1	402.333	377.216	423.041	391.7	361.674	360.008	450.925
WRDIMGc (Imagability for content words, mean)	426.9	422.707	419.167	438.245	409.5	384.208	387.123	455.06
WRDPOLc (Polysemy for content words, mean)	3.7	3.333	3.429	4.205	4.1	3.96	4.095	4.26

Index	iBT Reading Average[a]	iBT WR1	iBT WR2	iBT WR3	iBT Listening Average[b]	iBT WL1	iBT WL2	iBT WL3
RDFRE (Flesch Reading Ease)	50.3	51.644	34.44	64.427	58.6	53.935	57.604	63.731
RDFKGL (Flesch-Kincaid Grade Level)	10.6	11.158	12.033	8.628	9.5	10.858	9.138	8.813
RDL2 (Coh-Metrix L2 Readability)	16.6	10.355	19.14	19.622	13.9	10.21	17.675	13.16

[a]Average for the three reading passages, weighted by number of words. [b]Average for the three listening passages, weighted by number of words.

Table A5. Vocabulary Analysis of GEPT-A Reading Passages

Index	GEPT average[a]	GEPT R1	GEPT R2	GEPT R3	GEPT R4	GEPT R4b	GEPT R5	GEPT R6	GEPT R7	GEPT R7a	GEPT R7b	GEPT R7c
K1	79.7%	80.3%	82.2%	71.9%	78.2%	75.9%	83.3%	81.8%	79.3%	81.0%	78.6%	77.7%
K2	5.2%	5.3%	4.0%	4.7%	4.2%	5.6%	4.9%	4.9%	7.4%	7.6%	7.6%	6.5%
AWL	7.3%	3.5%	10.3%	9.8%	12.3%	12.4%	5.2%	5.5%	5.7%	6.9%	6.9%	4.2%
Off-list	7.2%	10.1%	2.4%	13.6%	5.3%	6.2%	6.6%	5.1%	7.0%	4.5%	6.9%	10.0%

Note. Due to rounding, totals may not exactly equal 100.0%.

[a]Average for the seven passages, weighted by number of words.

Table A6. Vocabulary Analysis of iBT Reading Passages

Index	iBT average[a]	iBT R1	iBT R2	iBT R3	iBT R4	iBT R5	iBT R6	iBT R7	iBT R8	iBT R9
K1	75.5%	79.1%	71.4%	77.4%	78.8%	76.6%	72.3%	73.6%	73.3%	76.9%
K2	6.5%	5.6%	6.5%	7.6%	3.7%	7.3%	11.2%	7.2%	4.0%	5.4%
AWL	7.7%	9.5%	8.0%	6.0%	3.7%	6.1%	4.1%	11.2%	13.3%	7.6%
Off-list	10.0%	5.8%	13.7%	9.1%	13.8%	8.3%	12.3%	7.9%	9.0%	10.2%

Note. Due to rounding, totals may not exactly equal 100.0%.

[a]Average for the nine passages, weighted by number of words.

Table A7. Vocabulary Analysis of Input Passages for GEPT-A Integrated Writing Tasks

Index	GEPT Average[a]	GEPT WR1	GEPT WR2
K1	70.2%	69.5%	70.9%
K2	12.5%	12.7%	12.2%
AWL	8.3%	8.4%	8.1%
Off-list	9.0%	9.4%	8.6%

Note. Due to rounding, totals may not exactly equal 100.0%.

[a]Average for the two passages, weighted by number of words.

Table A8. Vocabulary Analysis of Input Passages (Reading and Listening) for iBT Integrated Writing Tasks

Index	iBT Reading Average[a]	iBT WR1	iBT WR2	iBT WR3	iBT Listening Average[b]	iBT WL1	iBT WL2	iBT WL3
K1	79.2%	75.9%	72.8%	88.2%	82.9%	83.2%	81.2%	84.3%
K2	4.9%	6.0%	4.2%	4.7%	6.9%	7.3%	3.8%	9.8%
AWL	6.7%	4.9%	10.8%	4.4%	3.3%	1.5%	4.5%	3.6%
Off-list	9.0%	13.2%	11.5%	2.7%	5.8%	8.0%	7.6%	2.0%

Note. Due to rounding, totals may not exactly equal 100.0%.

[a] Average for the three reading passages, weighted by number of words.

[b] Average for the three listening passages, weighted by number of words.

Appendix B

Task Analysis Results for Individual Reading Items

Table B1. Aspects of Reading Assessed by Individual GEPT-A Items

Item	Passage	Aspect of reading assessed	Scope	Task format
1	GEPT 1 (Caravaggio)	Reading for specific details	Broad	Multiple choice
2	GEPT 1 (Caravaggio)	Reading for specific details[a]	Broad	Short answer
3	GEPT 1 (Caravaggio)	Reading for specific details	Moderate	Short answer
4	GEPT 1 (Caravaggio)	Reading for specific details	Narrow	Short answer
5	GEPT 2 (Value-added Teacher Ratings)	Reading for specific details[a]	Very narrow	Multiple choice
6	GEPT 2 (Value-added Teacher Ratings)	Reading for specific details	Narrow	Short answer
7	GEPT 2 (Value-added Teacher Ratings)	Reading for specific details	Narrow	Short answer
8	GEPT 2 (Value-added Teacher Ratings)	Reading for specific details[a]	Broad	Multiple choice
9	GEPT 2 (Value-added Teacher Ratings)	Reading for specific details	Narrow	Short answer
10	GEPT 3 (Hydrates)	Reading for the main idea	Very broad	Multiple choice
11	GEPT 3 (Hydrates)	Reading for specific details[a]	Very narrow	Short answer
12	GEPT 3 (Hydrates)	Identifying author purpose	Moderate	Short answer
13	GEPT 3 (Hydrates)	Reading for specific details	Moderate	Short answer
14	GEPT 3 (Hydrates)	Sensitivity to rhetorical organization	Very broad	Multiple choice
15	GEPT 4 (Brownfields)	Paraphrasing and/or summarizing	Narrow	Short answer

Item	Passage	Aspect of reading assessed	Scope	Task format
16	GEPT 4 (Brownfields)	Paraphrasing and/or summarizing	Narrow	Short answer
17	GEPT 4 (Brownfields)	Paraphrasing and/or summarizing	Narrow	Short answer
18	GEPT 4 (Brownfields)	Paraphrasing and/or summarizing[a]	Narrow	Short answer
19	GEPT 4 (Brownfields)	Paraphrasing and/or summarizing	Moderate	Short answer
20	GEPT 4 (Brownfields)	Paraphrasing and/or summarizing[a]	Moderate	Short answer
21	GEPT 5 (Hudson's Bay Company)	Skimming[b]	Moderate	Matching
22	GEPT 5 (Hudson's Bay Company)	Skimming[b]	Moderate	Matching
23	GEPT 5 (Hudson's Bay Company)	Skimming[b]	Moderate	Matching
24	GEPT 5 (Hudson's Bay Company)	Skimming[b]	Moderate	Matching
25	GEPT 5 (Hudson's Bay Company)	Skimming[b]	Moderate	Matching
26	GEPT 5 (Hudson's Bay Company)	Skimming[b]	Moderate	Matching
27	GEPT 6 (Victor the Wild Child)	Skimming[b]	Moderate	Matching
28	GEPT 6 (Victor the Wild Child)	Skimming[b]	Moderate	Matching
29	GEPT 6 (Victor the Wild Child)	Skimming[b]	Moderate	Matching
30	GEPT 6 (Victor the Wild Child)	Skimming[b]	Moderate	Matching
31	GEPT 6 (Victor the Wild Child)	Skimming[b]	Moderate	Matching
32	GEPT 6 (Victor the Wild Child)	Skimming[b]	Moderate	Matching
33	GEPT 7 (Three Historical Attractions)	Scanning	Very broad	Fixed multiple choice
34	GEPT 7 (Three Historical Attractions)	Scanning	Very broad	Fixed multiple choice

Item	Passage	Aspect of reading assessed	Scope	Task format
35	GEPT 7 (Three Historical Attractions)	Scanning	Very broad	Fixed multiple choice
36	GEPT 7 (Three Historical Attractions)	Scanning	Very broad	Fixed multiple choice
37	GEPT 7 (Three Historical Attractions)	Scanning	Very broad	Fixed multiple choice
38	GEPT 7 (Three Historical Attractions)	Scanning	Very broad	Fixed multiple choice
39	GEPT 7 (Three Historical Attractions)	Scanning	Very broad	Fixed multiple choice
40	GEPT 7 (Three Historical Attractions)	Scanning	Very broad	Fixed multiple choice

[a] LTTC considers these items to assess inferencing. [b] LTTC considers these items to assess both skimming and reading for the main idea.

Table B2. Aspects of Reading Assessed by Individual iBT Items (Practice Test 1)

Item	Passage	Aspect of reading assessed	Scope	Task format
1	iBT R1 (19th Century Politics in the United States)	Vocabulary knowledge/ determining the meaning of unfamiliar vocabulary from context	Narrow	Multiple choice
2	iBT R1 (19th Century Politics in the United States)	Reading for specific details	Very narrow	Multiple choice
3	iBT R1 (19th Century Politics in the United States)	Identifying author purpose	Very narrow	Multiple choice
4	iBT R1 (19th Century Politics in the United States)	Reading for specific details	Narrow	Multiple choice
5	iBT R1 (19th Century Politics in the United States)	Reading for specific details	Very narrow	Multiple choice
6	iBT R1 (19th Century Politics in the United States)	Vocabulary knowledge/determining the meaning of unfamiliar vocabulary from context	Very narrow	Multiple choice
7	iBT R1 (19th Century Politics in the United States)	Reading for specific details	Narrow	Multiple choice
8	iBT R1 (19th Century Politics in the United States)	Vocabulary knowledge/determining the meaning of unfamiliar vocabulary from context	Narrow	Multiple choice
9	iBT R1 (19th Century Politics in the United States)	Inferencing	Narrow	Multiple choice
10	iBT R1 (19th Century Politics in the United States)	Reading for specific details	Narrow	Multiple choice

Item	Passage	Aspect of reading assessed	Scope	Task format
11	iBT R1 (19th Century Politics in the United States)	Paraphrasing and/or summarizing	Narrow	Multiple choice
12	iBT R1 (19th Century Politics in the United States)	Sensitivity to rhetorical organization	Moderate	Multiple choice
13	iBT R1 (19th Century Politics in the United States)	Reading for major points	Very broad	Multiple response multiple choice
14	iBT R2 (The Expression of Emotions)	Vocabulary knowledge/ determining the meaning of unfamiliar vocabulary from context	Very narrow	Multiple choice
15	iBT R2 (The Expression of Emotions)	Identifying author purpose	Narrow	Multiple choice
16	iBT R2 (The Expression of Emotions)	Vocabulary knowledge/determining the meaning of unfamiliar vocabulary from context	Very narrow	Multiple choice
17	iBT R2 (The Expression of Emotions)	Sensitivity to cohesion	Narrow	Multiple choice
18	iBT R2 (The Expression of Emotions)	Reading for specific details	Very narrow	Multiple choice
19	iBT R2 (The Expression of Emotions)	Paraphrasing and/or summarizing	Very narrow	Multiple choice
20	iBT R2 (The Expression of Emotions)	Reading for specific details	Narrow	Multiple choice
21	iBT R2 (The Expression of Emotions)	Reading for specific details	Narrow	Multiple choice
22	iBT R2 (The Expression of Emotions)	Vocabulary knowledge/determining the meaning of unfamiliar vocabulary from context	Very narrow	Multiple choice
23	iBT R2 (The Expression of Emotions)	Vocabulary knowledge/determining the meaning of unfamiliar vocabulary from context	Very narrow	Multiple choice
24	iBT R2 (The Expression of Emotions)	Reading for specific details	Narrow	Multiple choice

Item	Passage	Aspect of reading assessed	Scope	Task format
25	iBT R2 (The Expression of Emotions)	Sensitivity to rhetorical organization	Narrow	Multiple choice
26	iBT R2 (The Expression of Emotions)	Reading for major points	Very broad	Multiple response multiple choice
27	iBT R3 (Geology and Landscape)	Reading for specific details	Narrow	Multiple choice
28	iBT R3 (Geology and Landscape)	Vocabulary knowledge/ determining the meaning of unfamiliar vocabulary from context	Very narrow	Multiple choice
29	iBT R3 (Geology and Landscape)	Paraphrasing and/or summarizing	Very narrow	Multiple choice
30	iBT R3 (Geology and Landscape)	Inferencing	Narrow	Multiple choice
31	iBT R3 (Geology and Landscape)	Vocabulary knowledge/determining the meaning of unfamiliar vocabulary from context	Narrow	Multiple choice
32	iBT R3 (Geology and Landscape)	Reading for specific details	Narrow	Multiple choice
33	iBT R3 (Geology and Landscape)	Identifying author purpose	Narrow	Multiple choice
34	iBT R3 (Geology and Landscape)	Vocabulary knowledge/ determining the meaning of unfamiliar vocabulary from context	Very narrow	Multiple choice
35	iBT R3 (Geology and Landscape)	Sensitivity to cohesion[a]	Very narrow	Multiple choice
36	iBT R3 (Geology and Landscape)	Reading for specific details	Narrow	Multiple choice
37	iBT R3 (Geology and Landscape)	Sensitivity to rhetorical organization	Moderate	Multiple choice
38	iBT R3 (Geology and Landscape)	Reading for specific details	Very broad	Classification

[a] Cohesion is normally taken as involving ties across sentence boundaries, following Halliday and Hasan (1976). In this case, however, the cohesive ties occur across independent clause boundaries, thereby bringing it into the realm of cohesion, rather than simply grammatical parsing.

Table B3. ***Aspects of Reading Assessed by Individual iBT Items (Practice Test 2)***

Item	Passage	Aspect of reading assessed	Scope	Task format
1	IBT R4 (Feeding Habits of East African Herbivores)	Vocabulary knowledge/determining the meaning of unfamiliar vocabulary from context	Narrow	Multiple choice
2	IBT R4 (Feeding Habits of East African Herbivores)	Vocabulary knowledge/determining the meaning of unfamiliar vocabulary from context	Very broad	Multiple choice
3	IBT R4 (Feeding Habits of East African Herbivores)	Reading for specific details	Moderate	Multiple choice
4	IBT R4 (Feeding Habits of East African Herbivores)	Vocabulary knowledge[a]	Very narrow	Multiple choice
5	IBT R4 (Feeding Habits of East African Herbivores)	Identifying author purpose	Very narrow	Multiple choice
6	IBT R4 (Feeding Habits of East African Herbivores)	Reading for specific details	Narrow	Multiple choice
7	IBT R4 (Feeding Habits of East African Herbivores)	Inferencing	Moderate	Multiple choice
8	IBT R4 (Feeding Habits of East African Herbivores)	Reading for specific details	Moderate	Multiple choice
9	IBT R4 (Feeding Habits of East African Herbivores)	Vocabulary knowledge/ determining the meaning of unfamiliar vocabulary from context	Narrow	Multiple choice
10	IBT R4 (Feeding Habits of East African Herbivores)	Vocabulary knowledge/determining the meaning of unfamiliar vocabulary from context	Very narrow	Multiple choice

Item	Passage	Aspect of reading assessed	Scope	Task format
11	IBT R4 (Feeding Habits of East African Herbivores)	Reading for specific details	Very narrow	Multiple choice
12	IBT R4 (Feeding Habits of East African Herbivores)	Reading for specific details	Broad	Multiple choice
13	IBT R4 (Feeding Habits of East African Herbivores)	Sensitivity to rhetorical organization	Narrow	Multiple choice
14	IBT R4 (Feeding Habits of East African Herbivores)	Reading for major points	Very broad	Multiple response multiple choice
15	iBT R5 (Loie Fuller)	Inferencing	Narrow	Multiple choice
16	iBT R5 (Loie Fuller)	Reading for specific details	Narrow	Multiple choice
17	iBT R5 (Loie Fuller)	Vocabulary knowledge[a]	Very narrow	Multiple choice
18	iBT R5 (Loie Fuller)	Paraphrasing and/or summarizing	Very narrow	Multiple choice
19	iBT R5 (Loie Fuller)	Vocabulary knowledge/determining the meaning of unfamiliar vocabulary from context	Very narrow	Multiple choice
20	iBT R5 (Loie Fuller)	Vocabulary knowledge/determining the meaning of unfamiliar vocabulary from context	Very narrow	Multiple choice
21	iBT R5 (Loie Fuller)	Reading for specific details	Moderate	Multiple choice
22	iBT R5 (Loie Fuller)	Reading for specific details	Narrow	Multiple choice
23	iBT R5 (Loie Fuller)	Identifying author purpose	Narrow	Multiple choice
24	iBT R5 (Loie Fuller)	Vocabulary knowledge/determining the meaning of unfamiliar vocabulary from context	Very narrow	Multiple choice

Item	Passage	Aspect of reading assessed	Scope	Task format
25	iBT R5 (Loie Fuller)	Reading for specific details	Narrow	Multiple choice
26	iBT R5 (Loie Fuller)	Reading for specific details	Broad	Multiple choice
27	iBT R5 (Loie Fuller)	Sensitivity to rhetorical organization	Moderate	Multiple choice
28	iBT R5 (Loie Fuller)	Reading for major points	Very broad	Multiple response multiple choice
29	iBT R6 (Green Icebergs)	Reading for specific details	Moderate	Multiple choice
30	iBT R6 (Green Icebergs)	Reading for specific details	Very narrow	Multiple choice
31	iBT R6 (Green Icebergs)	Paraphrasing and/or summarizing	Very narrow	Multiple choice
32	iBT R6 (Green Icebergs)	Vocabulary knowledge/determining the meaning of unfamiliar vocabulary from context	Narrow	Multiple choice
33	iBT R6 (Green Icebergs)	Reading for specific details	Very narrow	Multiple choice
34	iBT R6 (Green Icebergs)	Reading for specific details	Narrow	Multiple choice
35	iBT R6 (Green Icebergs)	Identifying author purpose	Broad	Multiple choice
36	iBT R6 (Green Icebergs)	Vocabulary knowledge/determining the meaning of unfamiliar vocabulary from context	Narrow	Multiple choice
37	iBT R6 (Green Icebergs)	Vocabulary knowledge/determining the meaning of unfamiliar vocabulary from context	Narrow	Multiple choice
38	iBT R6 (Green Icebergs)	Vocabulary knowledge/determining the meaning of unfamiliar vocabulary from context	Very narrow	Multiple choice
29	iBT R6 (Green Icebergs)	Reading for specific details	Very broad	Multiple choice
40	iBT R6 (Green Icebergs)	Inferencing	Broad	Multiple choice

Item	Passage	Aspect of reading assessed	Scope	Task format
41	iBT R6 (Green Icebergs)	Sensitivity to rhetorical organization	Moderate	Multiple choice
42	iBT R6 (Green Icebergs)	Reading for major points	Very broad	Multiple response multiple choice

[a] The meaning of the targeted word could not be inferred from context clues.

Table B4. Aspects of Reading Assessed by Individual iBT Items (Practice Test 3)

Item	Passage	Aspect of reading assessed	Scope	Task format
1	iBT R7 (Architecture)	Reading for specific details	Moderate	Multiple choice
2	iBT R7 (Architecture)	Vocabulary knowledge/determining the meaning of unfamiliar vocabulary from context	Narrow	Multiple choice
3	iBT R7 (Architecture)	Vocabulary knowledge/determining the meaning of unfamiliar vocabulary from context	Very narrow	Multiple choice
4	iBT R7 (Architecture)	Paraphrasing and/or summarizing	Very narrow	Multiple choice
5	iBT R7 (Architecture)	Vocabulary knowledge/determining the meaning of unfamiliar vocabulary from context	Very narrow	Multiple choice
6	iBT R7 (Architecture)	Vocabulary knowledge/determining the meaning of unfamiliar vocabulary from context	Narrow	Multiple choice
7	iBT R7 (Architecture)	Reading for specific details	Narrow	Multiple choice
8	iBT R7 (Architecture)	Inferencing	Very narrow	Multiple choice
9	iBT R7 (Architecture)	Inferencing	Moderate	Multiple choice

Item	Passage	Aspect of reading assessed	Scope	Task format
10	iBT R7 (Architecture)	Vocabulary knowledge/determining the meaning of unfamiliar vocabulary from context	Very narrow	Multiple choice
11	iBT R7 (Architecture)	Identifying author purpose	Narrow	Multiple choice
12	iBT R7 (Architecture)	Reading for specific details	Narrow	Multiple choice
13	iBT R7 (Architecture)	Sensitivity to rhetorical organization	Moderate	Multiple choice
14	iBT R7 (Architecture)	Reading for major points	Very broad	Multiple response multiple choice
15	iBT R8 (The Long-Term Stability of Ecosystems)	Vocabulary knowledge/ determining the meaning of unfamiliar vocabulary from context	Very narrow	Multiple choice
16	iBT R8 (The Long-Term Stability of Ecosystems)	Reading for specific details	Narrow	Multiple choice
17	iBT R8 (The Long-Term Stability of Ecosystems)	Reading for specific details	Very narrow	Multiple choice
18	iBT R8 (The Long-Term Stability of Ecosystems)	Reading for specific details	Narrow	Multiple choice
19	iBT R8 (The Long-Term Stability of Ecosystems)	Reading for specific details	Narrow	Multiple choice
20	iBT R8 (The Long-Term Stability of Ecosystems)	Reading for specific details	Moderate	Multiple choice
21	iBT R8 (The Long-Term Stability of Ecosystems)	Inferencing	Narrow	Multiple choice

Item	Passage	Aspect of reading assessed	Scope	Task format
22	iBT R8 (The Long-Term Stability of Ecosystems)	Vocabulary knowledge[a]	Very narrow	Multiple choice
23	iBT R8 (The Long-Term Stability of Ecosystems)	Identifying author purpose	Narrow	Multiple choice
24	iBT R8 (The Long-Term Stability of Ecosystems)	Vocabulary knowledge/determining the meaning of unfamiliar vocabulary from context	Narrow	Multiple choice
25	iBT R8 (The Long-Term Stability of Ecosms)	Paraphrasing and/or summarizing	Very narrow	Multiple choice
26	iBT R8 (The Long-Term Stability of Ecosystems)	Vocabulary knowledge/determining the meaning of unfamiliar vocabulary from context	Narrow	Multiple choice
27	iBT R8 (The Long-Term Stability of Ecosystems)	Sensitivity to rhetorical organization	Moderate	Multiple choice
28	iBT R8 (The Long-Term Stability of Ecosystems)	Reading for major points	Very broad	Multiple response multiple choice
29	iBT R9 (Depletion of the Ogallala Aquifer)	Reading for specific details	Very narrow	Multiple choice
30	iBT R9 (Depletion of the Ogallala Aquifer)	Reading for specific details	Narrow	Multiple choice
31	iBT R9 (Depletion of the Ogallala Aquifer)	Paraphrasing and/or summarizing	Very narrow	Multiple choice
32	iBT R9 (Depletion of the Ogallala Aquifer)	Vocabulary knowledge/determining the meaning of unfamiliar vocabulary from context	Narrow	Multiple choice

Item	Passage	Aspect of reading assessed	Scope	Task format
33	iBT R9 (Depletion of the Ogallala Aquifer)	Identifying author purpose	Narrow	Multiple choice
34	iBT R9 (Depletion of the Ogallala Aquifer)	Vocabulary knowledge[a]	Very narrow	Multiple choice
35	iBT R9 (Depletion of the Ogallala Aquifer)	Vocabulary knowledge/determining the meaning of unfamiliar vocabulary from context	Very narrow	Multiple choice
36	iBT R9 (Depletion of the Ogallala Aquifer)	Reading for specific details	Narrow	Multiple choice
37	iBT R9 (Depletion of the Ogallala Aquifer)	Reading for specific details	Narrow	Multiple choice
38	iBT R9 (Depletion of the Ogallala Aquifer)	Vocabulary knowledge[a]	Very narrow	Multiple choice
39	iBT R9 (Depletion of the Ogallala Aquifer)	Reading for specific details	Narrow	Multiple choice
40	iBT R9 (Depletion of the Ogallala Aquifer)	Reading for specific details	Narrow	Multiple choice
41	iBT R9 (Depletion of the Ogallala Aquifer)	Sensitivity to rhetorical organization	Broad	Multiple choice
42	iBT R9 (Depletion of the Ogallala Aquifer)	Reading for major points	Very broad	Multiple response multiple choice

[a] The meaning of the targeted word could not be inferred from context clues.

[b] No summarizing was involved on this item.

Appendix C

Additional Score and Item Analysis Results

Table C1. Descriptive Statistics for GEPT and iBT Scores (Raw Scores)

	GEPT	GEPT	GEPT	GEPT	iBT	iBT	iBT	iBT
	Reading1	Reading2	Reading	Writing	Reading	Writing	Listening	Speaking
Points possible	40	20	120[a]	5	30	30	30	30
Mean	23.0	11.6	69.4	2.6	24.9	24.0	24.4	22.2
Median	23	12	70.5	2.5	26	25	25	23
SD	6.7	4.5	21.9	0.5	4.2	3.5	4.3	3.0
Q	4.1	3.5	16.5	0.2	2.5	2.5	2.8	2.0
Skewness	-0.3	-0.2	-0.2	1.3	-1.9	-1.0	-1.5	-0.4
Kurtosis	0.2	-0.9	-0.6	1.6	5.4	1.7	3.2	1.2
Alpha	0.774	0.818	0.880	--	--	--	--	--
SEM	3.2	1.9	7.6	--	--	--	--	--

[a] Total GEPT reading score = Reading1 x 1.5 + Reading2 x 3.

Table C2. Descriptive Statistics for Raw Scores on Individual GEPT-A Passages (all test takers)

	R1	R2	R3	R4	R5	R6	R7
Mean	5.1	6.9	6.2	4.8	2.7	4.2	4.7
Median	5	7	6	5	2	5	5
SD	1.8	1.9	2.4	3.2	1.7	1.9	2.3
Q	1.0	1.0	1.5	2.5	1.5	1.5	2.0
Skewness	-0.4	-0.6	-0.6	0.1	0.3	-0.7	-0.4
Kurtosis	-0.4	0.3	0.0	-0.9	-0.9	-0.7	-0.7

Table C3. Item Analysis Results for GEPT-A Reading (Using Total GEPT-A Reading Scores)

Item	IF*	Discrimination	Item	IF*	Discrimination
R01	0.80	0.35	R21	0.40	0.32
R02	0.58	0.45	R22	0.68	0.35
R03	0.61	0.27	R23	0.33	0.16
R04	0.57	0.27	R24	0.49	0.47
R05	0.88	0.21	R25	0.42	0.35
R06	0.64	0.32	R26	0.40	0.56
R07	0.43	0.31	R27	0.81	0.48
R08	0.91	0.40	R28	0.66	0.43
R09	0.61	0.20	R29	0.70	0.47
R10	0.70	0.35	R30	0.70	0.45
R11	0.51	0.36	R31	0.65	0.46
R12	0.54	0.37	R32	0.66	0.50
R13	0.54	0.41	R33	0.71	0.47
R14	0.81	0.28	R34	0.61	0.32
R15	0.46	0.40	R35	0.50	0.42
R16	0.17	0.32	R36	0.57	0.46
R17	0.31	0.53	R37	0.64	0.41
R18	0.60	0.43	R38	0.57	0.44
R19	0.58	0.43	R39	0.55	0.50
R20	0.27	0.51	R40	0.60	0.57

Note. Discrimination was calculated as the correlation between items and total GEPT-A reading score, adjusted for autocorrelation.

Table C4. Item Analysis Results for GEPT-A Reading (Using Total Passage Scores)

Item	IF*	Discrimination	Item	IF*	Discrimination
R01	0.80	0.16	R21	0.40	0.45
R02	0.58	0.37	R22	0.68	0.26
R03	0.61	0.29	R23	0.33	0.23
R04	0.57	0.28	R24	0.49	0.39
R05	0.88	0.15	R25	0.42	0.40
R06	0.64	0.29	R26	0.40	0.48
R07	0.43	0.28	R27	0.81	0.50
R08	0.91	0.21	R28	0.66	0.50
R09	0.61	0.21	R29	0.70	0.49
R10	0.70	0.24	R30	0.70	0.52
R11	0.51	0.35	R31	0.65	0.51
R12	0.54	0.22	R32	0.66	0.52
R13	0.54	0.40	R33	0.71	0.39
R14	0.81	0.30	R34	0.61	0.39
R15	0.46	0.40	R35	0.50	0.44
R16	0.17	0.36	R36	0.57	0.51
R17	0.31	0.51	R37	0.64	0.33
R18	0.60	0.43	R38	0.57	0.44
R19	0.58	0.49	R39	0.55	0.50
R20	0.27	0.53	R40	0.60	0.51

Note. Discrimination was calculated as the correlation between items and total GEPT-A reading score, adjusted for autocorrelation.

Appendix D

Additional Results from Correlational Analyses

Table D1. ***Correlations (Pearson* r*) Among GEPT-A Subscores, iBT Section Scores, and Test Takers Survey Questions, with Significance and Sample Sizes***

		GEPT R1	GEPT R2	GEPT W	iBT R	iBT W	iBT L	iBT S	Diffic R	Diffic W	CntRel R	CntRel W	TskRel R	TskRel W
GEPT R1	*r*	1.000												
	Sig.	--												
	n	184												
GEPT R2	*r*	.701**	1.000											
	Sig.	.000	--											
	n	184	184											
GEPT W	*r*	.532**	.410**	1.000										
	Sig.	.000	.000	--										
	n	178	178	178										
iBT R	*r*	.457**	.414**	.338**	1.000									
	Sig.	.000	.000	.000	--									
	n	183	183	177	183									
iBT W	*r*	.425**	.316**	.385**	.545**	1.000								
	Sig.	.000	.000	.000	.000	--								
	n	183	183	177	183	183								

		GEPT R1	GEPT R2	GEPT W	iBT R	iBT W	iBT L	iBT S	Diffic R	Diffic W	CntRel R	CntRel W	TskRel R	TskRel W
iBT L	*r*	.460**	.332**	.317**	.620**	.591**	1.000							
	Sig.	.000	.000	.000	.000	.000	--							
	n	183	183	177	183	183	183							
iBT S	*r*	.388**	.270**	.359**	.328**	.603**	.552**	1.000						
	Sig.	.000	.000	.000	.000	.000	.000	--						
	n	183	183	177	183	183	183	183						
Diffic R	*r*	-.384**	-.336**	-.216**	-.181*	-.342**	-.190*	-.279**	1.000					
	Sig.	.000	.000	.004	.016	.000	.011	.000	--					
	n	177	177	173	176	176	176	176	177					
Diffic W	*r*	-.176*	-.052	-.177*	-.197**	-.223**	-.203**	-.197**	.337**	1.000				
	Sig.	.019	.494	.020	.009	.003	.007	.009	.000	--				
	n	177	177	173	176	176	176	176	177	177				
CntRel R	*r*	-.091	-.064	-.193*	-.024	-.105	-.070	-.063	.101	.125	1.000			
	Sig.	.225	.395	.011	.747	.165	.355	.402	.179	.097	--			
	n	178	178	173	177	177	177	177	177	177	178			
CntRel W	*r*	.146	.144	-.077	.000	-.032	-.067	.014	-.160*	.186*	.647**	1.000		
	Sig.	.052	.055	.313	.999	.673	.380	.854	.034	.013	.000	--		
	n	177	177	172	176	176	176	176	176	176	177	177		

		GEPT R1	GEPT R2	GEPT W	iBT R	iBT W	iBT L	iBT S	Diffic R	Diffic W	CntRel R	CntRel W	TskRel R	TskRel W
TskRel R	*r*	-.151*	-.259**	-.091	.046	-.114	.083	-.041	.130	-.007	.304**	.177*	1.000	
	Sig.	.045	.000	.235	.545	.132	.271	.585	.085	.924	.000	.019	--	
	n	177	177	172	176	176	176	176	176	176	177	176	177	
TskRel W	*r*	-.156*	-.168*	-.218**	-.116	-.079	-.127	-.048	.010	.241**	.373**	.532**	.481**	1.000
	Sig.	.038	.026	.004	.124	.297	.093	.527	.897	.001	.000	.000	.000	--
	n	177	177	172	176	176	176	176	176	176	177	177	176	177

**Correlation is significant at the 0.01 level (2-tailed). *Correlation is significant at the 0.05 level (2-tailed).

Table D2. Correlation Matrix for Variables Used in the Exploratory and Confirmatory Factor Analyses

	GEPT_Pssg1	GEPT_Pssg2	GEPT_Pssg3	GEPT_Pssg4	GEPT_Pssg5	GEPT_Pssg6	GEPT_Pssg7	GEPT_W	iBT_R	iBT_W
GEPT_Pssg1	1.000									
GEPT_Pssg2	.416**	1.000								
GEPT_Pssg3	.385**	.359**	1.000							
GEPT_Pssg4	.276**	.259**	.403**	1.000						
GEPT_Pssg5	.329**	.257**	.321**	.363**	1.000					
GEPT_Pssg6	.341**	.335**	.437**	.440**	.327**	1.000				
GEPT_Pssg7	.365**	.363**	.369**	.528**	.390**	.385**	1.000			
GEPT_W	.469**	.345**	.396**	.344**	.363**	.284**	.296**	1.000		
iBT_R	.377**	.449**	.328**	.235**	.350**	.330**	.283**	.338**	1.000	
iBT_W	.334**	.249**	.279**	.345**	.283**	.202**	.243**	.385**	.545**	1.000

**Correlation is significant at the 0.01 level (2-tailed). *Correlation is significant at the 0.05 level (2-tailed).

Acknowledgments

We would like to thank faculty members Dr. Sun-Young Shin at Indiana University and Dr. Becky Huang at the University of Texas, San Antonio and student research assistants Carol Celestine and Elizabeth Lee Tien from Nanyang Technological University, Singapore.

We would also like to thank the study participants from California State University, Fullerton; Indiana University, Bloomington; and the University of Texas, San Antonio; the study coordinators on these campuses (Yenlan Li at California State University, Fullerton, and Aurora Aguirre at the University of Texas, San Antonio); the study participants in Taiwan; and the study coordinators at Language Training and Testing Center, Taipei.

GEPT®

LTTC 全民英檢

The Language Training and Testing Center (LTTC)
No.170, Sec.2, Xinhai Rd., Daan Dist.,
Taipei City, 10663 Taiwan(R.O.C.)
Tel: +886-2-2735-2565
Email: geptgrants@lttc.ntu.edu.tw
Website: www.lttc.ntu.edu.tw