UNIVERSITY OF CALIFORNIA

Los Angeles

Investigating the Relationships between a Reading Test and Can-do Statements of Performance on Reading Tasks

A dissertation submitted in partial satisfaction of the
requirements for the degree Doctor of Philosophy
in Applied Linguistics

by

Hsin-min Liu

2014

ABSTRACT OF THE DISSERTATION

Investigating the Relationships between a Reading Test and
Can-do Statements of Performance on Reading Tasks

by

Hsin-min Liu
Doctor of Philosophy in Applied Linguistics
University of California, Los Angeles, 2014
Professor Lyle F. Bachman, Chair

One of the fundamental problems in language testing is the lack of adequate generalizability between what a test is measuring and what fulfills the learners' real world language use needs. It is important to recognize that no matter how precise a test measures a construct, if the way that a construct is defined and the way that test tasks are specified do not correspond to the domain of generalization in a meaningful way, test scores may never become adequate indicators of what learners can do with English in real life. This study investigated constructs and tasks of the General English Proficiency Test (GEPT) high-intermediate reading test to explicate the issues involved in generalizing test scores to non-test situations.

The study identified and demonstrated quantitatively and qualitatively the way and extent to which two distinct ways of conceptualizing reading constructs, the trait/curriculum-based and the task/domain-based approaches, could lead to divergent construct specifications, difficulty levels, item/text characteristics, and underlying factor structures, using approaches of expert judgments and confirmatory factor analysis. A total of 242 university students and six trained raters participated in the study. All the participants took the GEPT reading test and a task-based reading test developed based on the can-do statements in the Common European Framework of Reference (CEFR) and its designated Target Language Use (TLU) domains.

It was found that when items are more task-based and workplace specific, the less similarity they share with trait/curriculum based test items. The nature and the constituents of the reading comprehension construct shift. Not only do task-based and workplace specific items require a significantly higher amount of complex propositional content to be interpreted rather than recognized, they also demand a wider range and extent of language abilities (ideational, functional, and sociolinguistic) and strategic competence when making such interpretations in relation to context. Among all the combinations of language abilities, that of manipulative function and strategic demand appear to have the most effect on the complexity of reading construct. The ability to comprehend texts then is different from the ability to comprehend texts in context. The very nature of contextualization changes the nature and constituents of the comprehension construct.

Using Bachman and Palmer's (2010) Assessment Use Argument (AUA) framework, this study strongly suggests that the GEPT is not as meaningful or generalizable as the Language Training and Testing Center (LTTC) claims it is. GEPT test scores do not provide stakeholders with sufficient information about the ability to be assessed in the TLU domain, and the GEPT tasks do not have a sufficient degree of correspondence to the TLU tasks. Due to inadequate sampling of the target constructs and its task characteristics, GEPT test scores do not appear to generalize to performance in the target domain.

The dissertation of Hsin-min Liu is approved.

Peter M. Bentler

Noreen M. Webb

Lyle F. Bachman, Committee Chair

University of California, Los Angeles

2014

DEDICATION

To those who truly believed in me

Table of Contents

LIST OF TABLES

LIST OF FIGURES

ACKNOWLEDGEMENTS

I am indebted to my advisor, Professor Lyle Bachman, who has helped and supported me through the many years of my doctoral study. Without his guidance, understanding, patience, and encouragement, this dissertation would not have been possible. I will always be grateful for this wonderful academic journey I had with Lyle at UCLA. Becoming Lyle's student and being a part of LASSLAB and UCLA was my biggest dream. I want to thank Lyle for making my dream come true and helping me become a researcher and scholar in language assessment.

I want to express my sincere gratitude to Professor Peter Benter for his support, assistance, and insightful comments on this study, as well as for his great teaching over the years. I would also like to thank Professor Noreen Webb for serving on my committee and for providing her valuable comments and advice on this study. I am deeply thankful to all my teachers, especially Professors Steve Reise, John Schumann, Li Cai, and Frederick Erickson, and fellow LASSLAB members, Ick kyu Choi, Jonathan Schmidgall, Hongwen Cai, Youngsoon So, Yujie Jia, Hoky Min, Yasu Imao, and Huan Wang at UCLA. I feel truly fortunate to have had the opportunity to learn from and grow with them.

This dissertation was funded by a Small Grant for Doctoral Research in Second or Foreign Language Assessment. I must thank Educational Testing Service (ETS) for their financial support that assisted me in my data collection. I am also grateful to the Language Training and Testing Center (LTTC) in Taiwan for providing me with the test materials needed to complete this study. I am also obliged to all the National Taiwan University (NTU) students who participated in this study as well as to my UCLA fellow colleagues Kristen

Lindblom, kamaal Majeed, Laura Haupt, Carleen Velez, and Linda Jensen who helped me with item rating and coding. A special thank you goes to Chi-Hsiang Liu for his financial support throughout the past few years. Without all the above-mentioned help, achievement of this dissertation would have been very difficult, if not impossible.

Finally, I want to thank the people I met in North America, especially those in Ames, Iowa City, and Los Angeles. Their selfless belief in me and their truly free spirits have shaped me to become the person I am today. They have accompanied me and enriched my life through my spectacular 20s. I am forever grateful to all of you. Thank you.

VITA

MA	TESL/Applied Linguistics, Iowa State University, USA 2005
BA	English, Tamkang University, Taiwan 2001

Grants and Awards

Small Grant for Doctoral Research in Second or Foreign Language Assessment, ETS 2012

Study Abroad Scholarship, Ministry of Education, Taiwan 2011-2013

Nonresident Tuition Fellowship, Graduate College, UCLA 2006- 2008

Research Excellence Award, Graduate College, Iowa State University 2005

Employment

Language Instructor, UCLA Extension 2011

Test Rater, Test of Oral Proficiency, UCLA 2009-2010

Test Rater, ESL Composition Placement Test, UCLA 2009

Graduate Student Researcher, National Center for Research (CRESST), UCLA 2009

Teaching Assistant, Asian Languages Department, UCLA 2006 -2007

Test developer, Focus on Grammar, Pearson Longman 2004 -2005

Research assistant, Longman English Online, Pearson Longman 2002 - 2003

Publication

Chapelle, C., & Liu, H. M. (2007). Theory and Research: Investigation of "Authentic" CALL Tasks. In J. Egbert & E. Hanson-Smith (Eds.), *CALL environments: Research, Practice, and Critical Issues*, 2nd edition, (pp. 111-130). Alexandria, VA: TESOL Publications.

Presentations

Liu, H.M. (2011). *Building a construct validity argument for the GEPT high-intermediate reading test: A confirmatory factor analysis approach*. Paper presented at Language Testing Research Colloquium (LTRC), University of Michigan, June 23-25, 2011.

Liu, H.M. (2010). *Building a construct validity argument for the GEPT high-intermediate reading test: A confirmatory factor analysis approach*. Paper presented at Southern California Association of Language Assessment Research (SCALAR), UCLA, California, May 1st, 2010

Liu, H. M. (2007). *Assessing authenticity in language learning and assessment*. Paper presented at Southern California Association of Language Assessment Research (SCALAR), UCLA, California, May 12, 2007.

Liu, H. M. (2006). *An investigation of methods for assessing authenticity in computer-assisted language learning and assessment*. Paper presented at American Association for Applied Linguistics (AAAL) conference, Montreal, Canada, June 17-20, 2006.

Liu, H. M. (2003). *The authenticity of speaking tests in an online English course*. Paper presented at Midwest Association of Language Testers (MwALT) conference, Purdue University, West Lafayette, Indiana, October 18, 2003.

CHAPTER 1. INTRODUCTION

1.1 Context of the problem

Language learners spend a significant amount of their school life learning English worldwide, and naturally, they expect to see that the training they receive goes beyond the school setting and transcends into the kind of English ability useful for coping with real world challenges. In Taiwan for example, an average college graduate receives at least seven years of English education throughout middle school and college education. However, employers and educators in Taiwan to date continue to complain about college graduates' English ability, indicating that they generally lack of the ability to communicate in English (Wu and Wu, 2010). Apparently, what learners have learned and being tested on in schools are not well aligned with what they can do and are required to do with English in real life. This points out a fundamental problem, and an ongoing one, in language testing,—the lack of adequate generalizability between what a test is measuring and what fulfills the learners' real world language use needs. No matter how precise a test measures a construct, if the way that a construct is defined and the way that test tasks are specified do not correspond to the domain of generalization in a meaningful way, test scores may never become adequate indicators of what learners can do with English in real life.

The General English Proficiency Test (GEPT) is a battery of tests that were developed to fulfill the needs of educators and employers in Taiwan to make real life occupational and educational decisions. That is, scores from the GEPT are used by universities to make graduation decisions, and by potential employers to make selection decisions. Accordingly, the score inferences made available for public are written in forms of can-do statements related to some target language use (TLU) situations sampled from real life. For example, the can-do statements for the college level reading test states that "test takers who pass this skill level can read written messages, instruction manuals, newspapers, and magazines, and at work, he/she can read general documents, abstract, meeting minutes, and reports (LTTC, 2000)." The use of can-do statements implies that the construct to be measured corresponds closely to the language capacities underlying performances on tasks sampled from real life. This way of defining the construct is essential because it places the often absent TLU domain in the center of inference making. Such construct definition is also beneficial when promoting a shift to a more communicative orientation in English education (Kunnan and Wu, 2009). Recently, the developers of the GEPT claim to have linked scores on their test to an extended set of can-do descriptors, the Common European Framework of Reference (CEFR) (Wu and Wu, 2010). This again emphasizes their wants and the importance of relating the test score interpretations to real life capacities of English use in this local context.

However, instead of defining the test construct based on a needs analysis of the language that is required to perform those can-do tasks in the test takers' TLU domains, the construct was defined based on theory and curriculum. This is due to the fact that the GEPT was also designed to provide criterion reference of English proficiency corresponding to five levels of English learners in Taiwan, namely Junior high school, High school, College (non-English majors), Graduate school (College English majors), and professionals/specialists. To reflect the college level of English attainment, the high-intermediate level test was actually developed based on selected theories of language abilities, some commonly-used university level EFL textbooks, and consultation with English instructors from 16 universities (LTTC, 2000). Thus, in the reading test, three discrete language knowledge components are measured by three task types: the "sentence fill-in" section measures lexical and syntactical knowledge in sentences, the “gap-fill” section measures lexical and syntactical knowledge in context, and the “reading comprehension” section measures ability to understand main ideas, details, and inference made in passages. Each test item is designated to measure a specific language knowledge elements corresponding to the construct definition.

In other words, although the test developers are claiming that the GEPT measures performance on tasks in the CEFR can-do statements, these claims do not relate directly to how the test was developed, and nor have they provided evidence to support the equivalence between the two distinct constructs or performance outcomes. Thus, it is unclear in what way and to what extent the content covered in the test and the ability required to perform the multiple-choice test tasks corresponds to those of the can-do tasks. For instance, it is unclear in what way test tasks such as ‘sentence completion’ with a single word selection and ‘gap fill’ with word(s) and grammar point selection correspond to reading tasks in the TLU domain, such as reading texts for meaning and practical use to accomplish duties in social and work situations. For the ‘reading comprehension’ section of the GEPT, the tasks and texts included may differ considerably from the tasks and texts described in the can-do statements in at least three aspects: a) linguistic complexity of texts may vary, including length, text structure, amount of contextualization, amount of rhetorical variation, difficulty in vocabulary and grammar used, and amount and complexity of information provided, b) cognitive demand, such as the extent, amount, and type of language to be processed, and c) communicative demand, in that multiple-choice questions on the test differ from open-ended questions and completion of job tasks for specific purposes. Consequently, the extent to which test scores can be interpreted as indicators of future performance in the can-do tasks is in question because the inferential link between the test construct and the TLU domain is not well established.

Sampling performance based on theories of language knowledge is very different from sampling performance based on language use tasks taken from real life. The gap between these two approaches to defining the construct reflects the unsolved issue raised by the Taiwanese educators and employers; it shows the same discrepancy between school achievement and real life needs. After all, the ability required to solve discrete problems of language *per se* is not the same as the ability required to solve integrated problems of language use in specific contexts. The two distinctive ways of defining constructs are referred as the trait-focused approach and task-focused approach in language testing, and these two different approaches would eventually lead to different test specifications, test tasks, performance outcomes, and classification decisions (Bachman, 2007). They would also share divergent score-based interpretations given the nature of their fundamentally different ways of conceptualizing language ability. As a result, investigating and demonstrating that score interpretations are generalizable from one construct/domain to another is at the heart of justifying test use and score meaningfulness of the GEPT in this specific context.

A test battery that is intended for practical use relevant to the daily life of Taiwanese citizens and for social good must be held accountable to its stakeholders. That is, the test developers must demonstrate on what basis test scores are linked to performances on tasks in the CEFR can-do statements. If the test developers cannot demonstrate that an adequate correspondence between test scores and performance on can-do tasks is well established, the consequences of using the GEPT reading test for graduation and selection decisions may not be as beneficial to schools, employers, and general citizens as it claims to be. This study uses Bachman and Palmer's (2010) Assessment Use Argument (AUA), which is an argument-based approach to validation to guide its logic speculation. An AUA is a conceptual framework for linking test scores and score-based inferences to test use based on a series of claims and warrants. Thus, to support the claim that the GEPT test scores do reflect what learners can do with English in those TLU tasks, and can thus be used to predict their future performance in English, at least three warrants in the AUA need to be articulated and supported by empirical evidence. These warrants are:

1. The content and construct coverage of the GEPT reading test corresponds to the categories and descriptors of can-do statements in the CEFR.
2. Characteristics of the GEPT reading test tasks correspond to that of the reading tasks derived from the can-do statements.
3. The underlying construct of the GEPT reading test corresponds to that of the reading tasks developed based on can-do statements.

1.2 Purpose of the study

In order to justify these claims about the content and tasks in the GEPT and about the use of scores from the GEPT as predictors of what test takers can do with English in those designated tasks, the correspondences between the GEPT reading test and the can-do statements of performances in TLU reading tasks must be examined and determined. The study aims to investigate three levels of correspondences between the two domains: 1) content and construct coverage, 2) correspondence between the characteristics of test tasks and TLU tasks, and 3) the abilities that are engaged based on performance outcomes.

1.3 Research questions

In order to examine the correspondences between the GEPT reading test and the TLU reading tasks in depth, three questions are asked:

1. In what way and to what extent does the GEPT reading test construct and content cover the categories and descriptors of can-do statements in the CEFR?
2. In what way and to what extent do characteristics of GEPT reading test tasks correspond to those of the reading tasks derived from the can-do statements?
3. In what way and to what extent does the underlying construct of the GEPT reading test correspond to that of the reading tasks developed based on the can-do statements?

1.4 Significance of the study

As Bachman (1990) points out, "one of the most important and persistent problems in language testing is that of defining language ability in such a way that we can be sure that the test methods we use will elicit language test performance that is characteristic of language performance in non-test situations (p.9)." Different test developers tend to conceptualize and operationalize the construct, "language ability", differently even when the tests share exactly the same intended purposes and decisions to be made. This study will investigate the extent to which and how two distinct ways of conceptualizing constructs, trait-focused and task-focused approach, lead to divergent test development processes, performance outcomes, and classification decisions. In doing so, it will enrich our understanding of how and why discrepancies exist between what a test is measuring and what the test developers claim it measures in another domain or situation. The study will thus add to our current understanding about test generalizability and score inferences to non-test situations.

CHAPTER 2. LITERATURE REVIEW

In this chapter, current perspectives on conceptualizing language use and L2 construct are first described to reveal the multiple views held by different applied linguists and testing specialists. Then the differences between the two distinct conceptualizations to defining constructs, the trait-based approach and the task-based approach, are analyzed and summarized, and the fundamental problem of generalizing test scores across constructs and domains is discussed. Finally, Bachman and Palmer's (2010) 'Assessment Use Argument' framework is presented as a means of explicating the ways in which the discrepancy between the two approaches to defining the L2 constructs affects generalizability as well as meaningfulness of test score interpretations.

2.1 Current perspectives on language use and the L2 construct

Language use can be portrayed as a series of dynamic, situated, and individualized interactions among features of contexts, language abilities, and personal attributes of language users. The phenomenon is rather complex and it is still not fully understood by applied linguists and language testing specialists to date. A large set of contextual parameters and language components underlying the multidimensional L2 construct were identified, as in Bachman (1990) and Bachman and Palmer (2010); however, the nature and manner of how these parts interact as a whole to simultaneously change both the contextual facets and those who interact with them, as Chalhoub-Deville (2003) phrased it, remains unclear (Douglas, 2000; Bachman, 2007). Accordingly, the dynamism of such L2 interaction may not yet be fully characterized in the current mainstream testing practices. Although L2 construct theorists have not explicitly articulated and empirically demonstrated how the dynamic and localized aspects of the L2 construct might work in testing situations given their lack of generalizability across contexts, current perspectives in defining language use and the L2 construct have revealed salient characteristics and features that are critical for disentangling the construct of language use for testing purposes.

The first perspective as proposed by Lado (1961) and Carroll (1968) concentrates on the “linguistic aspect” of language use. They defined variables of language as comprising phonology, morphology, syntax, and lexicon. These linguistic elements could be tested separately by “discrete-point” tasks in the skills of reading, writing, listening, and speaking. The “sentence fill-in” section in the GEPT reading test, for example, would reflect this perspective of language use where discrete grammatical knowledge is tested in the skill of reading. Although this perspective was influential because it was one of the first approaches to clearly lay out the fundamental units of language use, it did not deal with the contextual factors that might affect the very nature of these linguistic elements. It was not until Canale and Swain (1980) that the L2 construct was expanded into a construct for communicative purposes. This second perspective as proposed by Canale and Swain (1980) and later Canale (1983) focuses on the “communicative aspect” of language use. They defined communicative competence as consisting of grammatical, sociolinguistic, discourse, and strategic competence. This perspective incorporated the sociolinguistic aspect of language use in addition to linguistic knowledge, distinguished the ability to process texts for meaning from linguistic features per se, and identified the ability to strategically manipulate different aspects of language knowledge as one of the competences underlying the L2 construct.

The third perspective as proposed by Bachman (1990) and Bachman and Palmer (1996,

2010) perceived language use as a function of the “interactions” among features of contexts, language abilities, and personal attributes. That is, performance in using a language is an effect of both an individual’s language ability and personal attributes and of the characteristics of the context. This perspective, thus, described language use as involving two kinds of interactions: i) those among attributes of an individual language user; that is, when using language, the individual user’s topical knowledge, personal attributes, affective schemata, and cognitive strategies are engaged and interact with each other in addition to their language ability, and ii) those between the language user and the characteristics of a language use situation; situational features as outlined in their framework of task characteristics include setting, rubric, language input, expected response, and the relationship between input and expected response. Only when these interactions are involved can language ability be fully engaged in actual situational uses. Language ability, built on the Canale and Swain (1980) model, was defined as comprising two components: i) language knowledge, which consisted of organizational (grammatical and textual knowledge) and pragmatic knowledge (functional and sociolinguistic knowledge), and ii) strategic competence, which involved metacognitive strategies that manage the ways in which language users set goals, appraise, and plan action sequences in specific language use situations. This perspective significantly broadened the concept of language use and the L2 construct by clearly articulating the variables of context and personal attributes that were involved in L2 communication and interaction in addition to the communicative language abilities, offering the most comprehensive conceptualization of language ability at the time (Purpura, 2007).

The fourth perspective, the direct, performance-based or task-based approach, apart from the above ability-focused approaches, emphasized the qualities of "performance outcomes" elicited in simulated real life tasks (Bachman, 1990, 2002a, 2002b, 2007; Long & Norris, 2000; Brown et al., 2002; Jones, 1985; Clark, 1975). This approach typically requires language users to "engage in some sort of behavior which stimulates, with as much fidelity as possible, goal-oriented target language use outside the language test situation. Performances on these tasks are then evaluated according to pre-determined, real-world criterion elements (i.e., task processes and outcomes) and criterion levels (i.e., authentic standards related to task success)" (Brown et al, 2002:10). From this viewpoint, language use is a function of performance on the task itself rather than an effect of ability and of context underlies the performance outcome. Drawing on research in education, second language acquisition, and language teaching, this perspective produces inferences about what language users can perform or can do with language on tasks sampled from the domain of generalization. Therefore, this perspective does not have a central organized model to define its elements and constituents; instead, it varies depending on the kind of tasks being sampled from the target domain. The more precise the sampling and the characterization of tasks, the more accurate one can expect the prediction of future performance on real-world tasks. It seeks to establish a direct representation of real-world language use in testing situations without making an indirect inference about real-world performances based on a currently accepted theory of language ability.

2.2 Differences between the two distinct constructs: trait-based vs. task-based

L2 researchers and theorists view the same phenomenon, second language use, from different perspectives, and accordingly, practitioners and language testers use of these different viewpoints to put together a test suitable for their intended needs. While the dynamism of L2 interaction itself has not been fully characterized and understood, current perspectives to defining language use and the L2 construct can be divided into two main conceptualizations: the trait-based approach and the task-based approach, each approach deriving their approaches from very different sets of values and assumptions (Bachman, 1990, 2002a, 2002b, 2007; Skehan, 1998; Chapelle, 1998; Upshur, 1979; Norris et al, 1998).

The trait-based approach focuses on expanding the definition of the "underlying language capacity" that enables a language user to perform, and the construct definition has evolved from referring to mere linguistics theory, to communicative competence, to the full range of interactions among features of context, language abilities, and personal attributes. In this approach, the test is developed on the basis of a sampling of known language abilities or knowledge of interest to test developers and users, and thus, test scores are interpreted as test takers possessing specific and identifiable sets of language abilities or knowledge. This approach, however, is limited in that it tends to ignore the specifications and manifestations of contextual features in developing tests, and thus is subject to lack of sufficient sampling of test takers' other attributes that are also critical in using language in the target domain. This may lead this approach's test result to become a function of mere language knowledge but not a function of real language use. More specifically, some combinations of ability elements and their interaction effect are simply not definable or understood yet by applied linguists, and the trait-based approach can thus be limited and isolated to the extent that it fails to capture the complexity of language use in real life. The strength of this approach is that since it consistently samples from known theories or elements of language abilities, it offers concrete and specific inferences about the ability tested, advantages that the task-based approach cannot offer.

The task-based approach, in contrast, instead of depending on a model of language knowledge or ability, emphasizes what a language user "can do" or "can perform" with the L2 in real life based on the direct sampling of language use performances in real life simulated tasks. In this approach, the test is developed based on a needs analysis of representative tasks in the domain of interest, and test scores are interpreted as to what degree test takers can accomplish these tasks as required in real-life situations. Accordingly, such a test is primarily used to predict future performance in the target language use domain. This approach focuses on the recreation of the contextual features in test tasks, and often engages a wider range of other abilities (topical, affective schemata) in addition to language ability. While such engagement may reflect more of the complexity of language use in real life, specific language ability involved in this situation is "inextricably meshed with other abilities (topical, affective schemata) and with the test method used to elicit language" (Bachman, 2002:5). Therefore, this approach is limited in that the language abilities measured are not specific, concrete, and definable, and the abilities tested also vary when task features change from context to context.

In conclusion, sampling performances based on language knowledge involve very different sets and aspects of abilities than do sampling performances based on real-life simulating tasks. The former tends to focus on specific language abilities to be assessed, and neglects other aspects of language abilities and personal attributes that would also be engaged in language use in context, while the latter tends to engage a more robust range, but indefinable sets of language abilities. This gap results because there is no direct correspondence between language form and language function, and between theories of language and the ability required by tasks. Language form and knowledge must be adapted to fit specific functions of use in numerous manners, and one may not assume that test takers have the capacity to transfer what they know from context to context without supporting it with evidence. Table 2.1 summarizes the differences between the two distinct constructs from the various aspects discussed earlier.

Table 2.1: Distinction between the two constructs

Aspects	Trait/ability-based construct	Performance/Task-based construct
Interpretation	Has X ability	Can do Y
Use	Infer ability	Predict future performance
Definition	Underlying ability or capacity	Performance on tasks
Reference	Theories	Real life needs analysis
Domain of sampling	Theories of language ability	Real life domains or contexts
Elements	Grammar, vocabulary, cohesion, rhetorical organization, functional and sociolinguistic appropriateness, strategic competence	Tasks (settings, requirements, criterion for evaluation, etc)
Nature of engagement	Specific aspects of language ability	Language ability inextricably meshed with other abilities (topical, affective schemata) and with the method
Consistency	Trait factors	Contextual factors
Inconsistency	The ability engaged does not equate to that required in real life tasks	The ability engaged varies from task to task
Main criticism	Not generalizable to the TLU domain	No specificity on ability measured
Task type	Selected response	Open-ended and integrated response

2.3 Generalizability and meaningfulness in the Assessment Use Argument

Argument-based approaches to validation have provided for the field of educational measurement a framework for evaluating and justifying intended score interpretations (Kane, 1992; Kane, Crooks, & Cohen, 1999; Kane 2001; Mislevy, Steinberg, & Almond, 2003). These approaches, however, have not yet developed a coherent set of guidance for linking test scores and score-based inferences to test use. Bachman (2005) builds on these argument-based approaches and proposes an extended conceptual framework called the ‘assessment use argument’ (AUA) that not only links performance to score interpretation, but also links score interpretation to test use. Bachman and Palmer (2010) clearly outline an AUA as a series of data-claim inferential links based on Toulmin’s (2003) structure of practical reasoning, and incorporate critical measurement qualities into each step of inference linking to support the outcomes of claims. By utilizing an AUA, a particular use of the test can be justified based on claims and warrants stated and backings collected.

Therefore, when test developers claim that a test constructed based on traits can predict performances on tasks in real-life domains, this poses serious problems of test score interpretation meaningfulness and generalizability. Meaningfulness, as indicated in Bachman and Palmer (2010), refers to “the extent to which a given assessment record (i) provides stakeholders with information about the ability to be assessed, and (ii) conveys this information in terms that they can understand and relate to (p.114),” and generalizability refers to “the degree of correspondence between a given language assessment task and a target language use task in their task characteristics (p.117).” Accordingly, in the case of GEPT, the information provided to stakeholders, which is presented in the format of can-do statements, does not match with the ability actually assessed in the GEPT and no empirical evidence is presented to support a correspondence between the GEPT test tasks and tasks in the can-do statement. As discussed earlier, the trait-based approach to construct definition is very different from the task-based approach, and consequently, the two distinctive ways of conceptualizing constructs would lead to different test specifications, test task designs, performance outcomes, score-based interpretations, and classification decisions (Bachman, 2007). Unless the test developers demonstrate on what basis test scores are linked to performances on tasks in those can-do statements, test scores may not be considered as adequate indicators for educators and employers in Taiwan to make relevant occupational and educational decisions. In order to articulate the magnitude of the issue one aspect at a time, claims and warrants in the AUA related to the case of GEPT that must be articulated and investigated are listed as follow:

Claim 3: The interpretations about the ability to be assessed are **meaningful** because

- Warrant A1: The definition of the construct is based on a frame of reference.
    - ✧ Trait-based approach: The constructs to be assessed include knowledge of lexicon, syntax, cohesion, and comprehension of main ideas, details, and inferences in reading passages. These constructs are based on selected theories of reading ability.
    - ✧ Task-based approach: The constructs to be assessed are performances on target language use tasks (memos generated, priorities determined, new knowledge acquired, instructions followed, opinions voiced). These construct definitions are based on needs analyses of reading tasks in the target language use domains.
- Warrant A2: The assessment task specifications clearly specify the conditions under which we will observe or elicit performance from which we can draw inferences about the construct we intend to assess.
    - ✧ Trait-based approach: The assessment task specifications clearly specify that the test takers will choose the most correct answer from the multiple options

given to complete certain random sentences and passages, and to answer questions about the content of other passages. The choices offered in the multiple choice questions focus on knowledge of lexicon, syntax, cohesion, and comprehension.

  - ✧ Task-based approach: The assessment task specifications clearly specify that the test takers will read magazine articles, business documents, and academic texts taken from the TLU domains and complete tasks to show that they can determine priorities, generate memos, maintain customer relationships, follow instructions, acquire content knowledge, and provide personal opinions about the readings.

- Warrant A4: The procedures for producing an assessment record focus on those aspects of the performance that are relevant to the construct we intend to assess.
  - ✧ Trait-based approach: The machine scored answer key focuses on knowledge of lexicon, syntax, cohesion, and comprehension.
  - ✧ Task-based approach: The scoring key and procedures for using the key focus on the criteria (accuracy, clarity, etc) used in the TLU domain.
- Warrant A5: Assessment tasks engage the ability defined in the construct definition.
  - ✧ Trait-based approach: Assessment tasks engage the knowledge of lexicon, syntax, cohesion, and comprehension.
  - ✧ Task-based approach: Assessment tasks engage the kind of performances required in the TLU domains.
- Warrant A6: Assessment records can be interpreted as indicators of the ability to be assessed.

    - ✧ Trait-based approach: Assessment records can be interpreted as indicators of test takers' having the knowledge of lexicon, syntax, cohesion, and comprehension.
    - ✧ Task-based approach: Assessment records can be interpreted as indicators of test takers' future performance in the TLU domain.

Claim 3: The interpretations about the ability to be assessed are **generalizable** to the TLU domain in which the decision is to be made

- Warrant C1: The characteristics of the setting, rubric, input, expected response, and relationship between input and expected response of the assessment tasks correspond closely to those of TLU tasks.
    - ✧ Trait-based approach: The characteristics of the tasks correspond closely to those of tasks of syntax, lexicon, cohesion, and comprehension in the college reading curriculum
    - ✧ Task-based approach: The characteristics of the tasks correspond closely to those of tasks of reading magazine articles, business documents, and academic texts in personal, work, and academic domains.
- Warrant C2: The criteria and procedures for recording the responses to the assessment tasks correspond closely to those that are typically used by language users in assessing performance in TLU tasks.
    - ✧ Trait-based approach: The criteria and procedures for evaluating the responses to the tasks correspond closely to those that are typically used by English instructors in colleges and universities.

✧ Task-based approach: The criteria and procedures for evaluating the responses to the tasks correspond closely to those that are typically used by friends, supervisors, and teachers in assessing performance in reading magazine articles, business documents, and academic texts.

To investigate the problem of correspondence between what a test is measuring and what is required to perform simulated real-life tasks, one must articulate all these warrants of meaningfulness and generalizability and support them with evidence. The research inevitably includes examining not only the correspondence of contextual features across domains, but also the ability required to perform tasks across domains. After all, language use is an effect of both language abilities and contextual features, and collecting evidence to support these aspects of correspondence between test performance and language use in the TLU domain is at heart of resolving the issue of generalizing test scores to non-test situations.

CHAPTER 3. METHOD

3.1 Research approaches

This study employed a mixed method approach, combining both qualitative and quantitative methods, to investigate the relationship between the GEPT reading test and its domain of generalization—real life performance in the format of can-do statements. In order to create a parallel comparison between performances across the two domains, a task-based test (TBT) was developed to elicit test takers' performance on tasks sampled from the designated TLU domain. In other words, the TBT served as a surrogate for the TLU domain and was used for the purpose of comparison with the GEPT reading test. The CEFR was used in this study because it provided a refined specification of the TLU domain by listing a series of can-do statements, documenting the type of reading construct, context, and task content needed to achieve the target performance in real life. While the can-do statements may present problems of specificity, in terms of score interpretations, they nevertheless provide a valuable tool for refining the real life domain for the purpose of the TBT development and for the cross-test comparison in this study.

The correspondence between the GEPT reading test and the TBT was examined at the level of score interpretation from three perspectives, a) the extent of the content/construct coverage of the CEFR, b) the correspondence between test task and TLU task characteristics, and c) the correspondence of the underlying construct based on actual performance outcomes, using approaches of expert judgment, task characteristic analysis, and factor analysis. The three aspects in turn correspond to the three warrants and the three research questions under investigation. Both the GEPT and the TBT were administered to a common set of examinees, and the test takers' response was collected and analyzed using an ex-post-facto correlational approach rather than experimental approach because only the correlational and covariance structure of the test response data were of concern, and no particular conditions or treatments were required to superimpose on the examinees and nor were they relevant to the variables under investigation in the study.

3.2 Participants

The high-intermediate GEPT test was designed to represent a proficiency level that was equivalent to a university graduate whose major was not English, and so the target population to which the study was intended to generalize was college students in Taiwan. Although the population of actual test takers might also contain high school students, graduate students, and adults from work since anyone with the age of 12 and above could register to take the test, they were not included in this study because they did not represent the major target for which the test was designed. A total of 242 paid college students were recruited from a wide variety of departments and with different levels of English proficiency from a university in northern Taiwan. The procedures for the selection and participation of research participants met the requirement of IRB review procedures. Their characteristics matched the target population as specified in Table 3.1.

Table 3.1: Characteristics of the total population from a single administration (LTTC, 2008)

Characteristics	Proportion
Scores/ability	15% 100-120, 35% 80-99, 34% 60-79, 14% 40-59, 2% 0-39 Mean= 78.97/120, standard deviation= 18.95, no range restriction
Region	57% north, 24% central, 17% south, 1% eastern
Gender	42% male and 58% female
Age	42% age 15-19, 37% age 20-24, others 21%
Education level	57% university, 25% high school, 8% graduate school, others 11%
Major	English 10% , science and engineering 18%, medical 9%, business 8%, others (humanities, law, education, psychology) 18%
Occupation	student 81%, nonstudent 19%
Purpose	58% self evaluation, 28% for education evaluation purposes, 9% education advancement, 3% job requirement, 1% job promotion

3.3 Raters

Six paid raters from the Departments of Applied Linguistics and English and the Writing Program at UCLA were recruited to analyze the CEFR levels and task characteristics of the GEPT and TBT items/tasks. Four were doctoral and master's degree students in applied linguistics and English, and the other two were lecturers teaching in the Writing Program who held applied linguistics degrees. They were experienced ESL/EFL teachers with many years of experience, both domestically and overseas. All the raters were trained according to the procedures recommended in the CEFR Manual (Council of Europe, 2003) prior to their coding, analysis, and classification.

3.4 Materials

Four sets of materials were used in the study: 1) a GEPT high-intermediate reading test paper was requested from the Language Training and Testing Center (LTTC) in Taiwan, 2) the CEFR and the manual of relating language tests to the CEFR (Council of Europe, 2003) were used for rater training and level coding, 3) a task-based reading test (TBT) was developed based on the CEFR and the TLU content experts, and 4) a task characteristics analysis instrument was developed based on the CEFR (Council of Europe, 2001), Dutch Grid (Alderson et al, 2006), Carr (2003), and Bachman and Palmer (2010).

3.4.1 The GEPT reading test

The high-intermediate reading test comprises three sections designed to measure three reading abilities: 1) the "sentence fill-in" section which is intended to measure lexical and syntactical knowledge in sentences, 2) the "gap-fill" section which is intended to measure lexical and syntactical knowledge in paragraphs, and 3) the "reading comprehension" section which is intended to measure the ability to understand main ideas, details, and inference made in passages. Each test item is designed to measure specific area of language knowledge corresponding to the test specification as shown in Table 3.2 below.

Table 3.2: GEPT Test specification

Sec	Type of task	# of item	Traits to be measured (# of item)	Amount of input
I	Sentence fill-in	10	Lexis (5) Syntax (5)	10 sentences
II	Gap-fill	15	Global Lexis (8) Local Lexis (2) Global syntax (5)	2 paragraphs
III	Reading comprehension	20	Main idea (5) Details (8) Inference (7)	4 passages (290 words/ average)

Both the sentence fill-in and the gap-fill sections are designed to measure lexical and syntactical knowledge, yet they differ in the amount of input to be processed. The sentence fill-in section includes ten sentences, each with one blank to fill in, which the gap-fill section contains two paragraphs with a total of fifteen gaps to complete. The reading comprehension section consists of a graphic task and four passages, and each passage contains 166 to 410 words in length. In general, the test has more items measuring lexical and syntactical knowledge than reading comprehension. All 45 items are dichotomously-scored, four-option multiple-choice items. Test takers were given 50 minutes to complete the test.

3.4.2 The CEFR and the Manual

The Common European Framework of Reference (CEFR) (Council of Europe, 2001) is a framework that was developed to describe the language use of language learners for communication purposes in a wide range of real life situations. It adopts an action-oriented approach to describing language performance, stressing the role of learners as "'social agents who have tasks to accomplish in a given set of circumstances in a specific environment and within a particular field of actions (p.9)." The CEFR documents general characteristics of the locations, persons, events, and texts in four main domains, personal, public, occupational, and educational. It also specifies a large set of can-do descriptors by skills, reception, production, interaction, at six levels of ability (A1-C2). The can-do descriptors essentially state actions of particular language uses that are performed in specific language tasks and situations. The high-intermediate GEPT, as claimed by its developers, has been linked to the CEFR B2 level in terms of content and level (Wu and Wu, 2010). Therefore, those who pass this level of GEPT reading test, as indicated in several places in the CEFR (Council of Europe, 2000), "can read with a large degree of independence, adapting style and speed of reading to different texts and purposes, and using appropriate reference sources selectively. Has a broad active reading vocabulary, but may experience some difficulty with low frequency idioms" (p.69), "can understand the main ideas of complex text on both concrete and abstract topics, including technical discussions in his/her field of specialization" (p.24), and "can read articles and reports concerned with contemporary literary prose in which the writers adopt particular stances" (p.27). The characteristics of reading tasks and their contexts and the can-do statements of reading at B2 level outlined in the CEFR were being used as a basis to define the construct of the task-based test. The CEFR descriptors and the Manual for relating language tests to the CEFR were used for rater training and level coding. The manual documents the standardized procedures and methods for training and coding. The rater training followed the procedures of familiarization and standardization in the *Relating language examinations to the Common European Framework of Reference for languages: learning, teaching, assessment. A Manual* (Council of Europe, 2003) prior to coding, analysis, and classification.

3.4.3 The Task-Based test (TBT)

A task-based test (TBT) was developed to collect information about learners' ability to use English in specific reading tasks sampled from the reading specifications/descriptors and language use domains outlined in the CEFR. Not only have the test developers claimed that the GEPT is linked to the CEFR, but also that the CEFR provides specifications intended d to be useful for defining a test construct more directly related to real life. Therefore, the TBT was specifically designed based on the CEFR B2 level descriptors/specifications of reading comprehension. In the CEFR manual, Table A2, "Salient Characteristics of Reading" (see the Appendix A), was particularly useful in developing the TBT reading comprehension items. Since the CEFR can-do statements served only as a reference for test development and often lacked specificity, a needs analysis of the TLU domain was conducted. A group of content experts from the designated social and workplace domains in real life was consulted. The samples of texts and tasks that these content experts provided, based on their daily work, became the reading tasks in the TBT social test (ST) and work test (WT).

The content experts included three student affairs officers and two undergraduate students at UCLA and a series of commentaries posted in online forums related to the two social reading comprehension (SRC) articles in the TBT. Two work reading comprehension (WRC) texts were selected from a pool of texts provided by the student affairs officers. Two SRC texts were chosen based on their relevancy to the life of Taiwanese college students and the characteristics detailed in the B2 level descriptors. Two TBT social tasks were taken from the commentaries in the online forums, and one was sampled from a recorded discussion between the two UCLA undergraduates after they read the two social articles. All three ST dialogues were derived from actual comments made in real life contexts. Since no direct tasks were associated with the selected WRC texts, the WT tasks were developed based on the critical work task characteristics described by the student affairs officers. The officers reported that they dealt mostly with problems and concerns raised by their supervisors and college students and that they had to consult a significant amount of general documents in order to resolve the questions asked in office emails and messages. Departmental clerical work on a college campus was selected because the work content was the most relevant and familiar to college undergraduates compared to other business settings.

Table 3.3: TBT Test specification (based on the CEFR and the TLU content experts)

Sec	Type of task	# of pts	Traits to be measured	Amount of input
I. a	Social reading comprehension (SRC)	20	Retrieve information Interpret meaning	Harvard text (415 words) Facebook text (726 words)
I. b	Social tasks (ST)	9	Respond to prompts Synthesize and report	Harvard text Facebook text 3 dialogues (50 words/dialogue)
II. a	Workplace reading comprehension (WRC)	13	Retrieve information Interpret meaning	GEOG text (619 words) SPUR text (941 words)
II. b	Workplace tasks (WT)	12	Respond to prompts Paraphrase and advise	GEOG text SPUR text 3 emails/message (56 words/prompt)

As can be seen in Table 3.3, the TBT sampled tasks from two domains, the personal and occupational domain, out of four domains outlined in the CEFR. Two aspects of the TBT construct were highlighted and distinguished: one was to be able to comprehend texts in specific domains and the other was to be able to use the information to solve problems or perform actions required to carry out duties in the designated situations. The TBT tasks asked test takers to a) respond to a supervisor and student customers to solve a specific problem based on a series of business emails and general work regulations in a work situation and b) respond to comments of a news story and a magazine article to friends or online in a personal social setting. The TBT comprehension item types included fill-in-blanks using content words and short answer questions, whereas the TBT task types included brief message/emails writing and dialogue completion.

3.4.4 Task characteristic specification

An instrument for analyzing the content of the GEPT and TBT was designed based on the literature on content/task analyses. A list of variables was specifically defined based on the CEFR manual descriptors (Council of Europe, 2003), the Dutch Grid (Alderson et al, 2006), Carr, (2003) and Bachman and Palmer's (2010) frameworks of task characteristic and components of language ability. Features identified include a variety of variables relating to linguistics complexity (syntax and lexicon), types of operation (retrieve, interpret, analysis), language functions (manipulative and ideational), text features (propositional complexity, rhetorical features and organization, abstractness, explicitness, sociocultural specificity), and expected response (strategic demands, response type, and scoring criteria). See Appendix B for a detailed specification of the instrument.

3.5 Procedures

The TBT was developed in spring 2012 based on the CEFR descriptors and consultation with TLU content experts. Subsequently, trials of the TBT were held with small groups of EFL college examinees and native speakers at a particular university in Taiwan throughout summer 2012. The test items and scoring methods underwent a series of revisions based on the examinees' responses and feedback. During this time, two GEPT High-Intermediate reading tests were requested from and provided by the LTTC, and only the version previously administered (instead of the one used for research purposes) was chosen for task characteristics analysis and data collection. Next, both tests were administered to 242 college examinees in fall 2012. Each student was given a maximum of three hours to complete the three tests: the GEPT, TBT-social, and TBT-work. Finally, six raters from the Departments of Applied Linguistics and English and the Writing Program at UCLA were recruited to analyze the CEFR levels and task characteristics of the GEPT and TBT items/tasks. They were trained according to the CEFR manual before performing coding, analysis, and classification.

3.6 Data Analysis

The correspondences between the GEPT reading test and the TB test were investigated from multiple perspectives. Hence, the data were analyzed using both qualitative and quantitative approaches.

3.6.1 Qualitative analysis

To answer the first research question, the extent to which the GEPT relates to CEFR B2 reading descriptors was investigated using expert judgments to qualitatively code and map the construct and content of the GEPT test, based on the categories and descriptors in the CEFR. To answer the second research question, the task characteristics of the two tests were analyzed by the raters based on the variables identified in the Appendix B. The task characteristics were manually classified, counted, rated, and documented. The results were qualitatively compared across tasks and tests.

3.6.2 Statistical analysis

Multiple statistical analyses were conducted to examine the different types and strengths of the relationships between the outcomes of the two tests. The underlying factor structures of both tests were inspected using confirmatory factor analysis. The test scores collected from both tests were analyzed following steps a-c below, using the Statistical Package for Social Sciences for Windows Release 15.0 (SPSS Inc., 2006) and EQS Version 6.1 (Bentler and Wu, 2007).

a. Grouping of items

Item level data were not used in the SEM analyses because these were not of central concern in this study. Instead, items measuring the same sub-construct in each test were bundled together and form the base unit for further statistical analyses. Table 3.4 lists all the variables to be modeled. For the GEPT reading test, ten variables were identified: 1) “lexicon” and 2) “syntax” as measured in the “sentence fill-in” section; 3) “cohesion”, 4) “gsyntax” (syntax), and 5) “glexicon” (lexicon) as measured in the “gap-fill” section, and 6) “graph”; 7) “B1 retrieve”, 8) “B1 understand”, 9) “B2 retrieve”, and 10) “B2 understand” as measured in the “reading comprehension” section. For the TBT, twelve variables were identified: four variables in social reading comprehension, four in work reading comprehension, three in social task completion, and three in work task completion.

Table 3.4: Variables to be modeled

Test	GEPT (G)		TBT			
Section	Linguistics knowledge (GLK)	Reading comprehension (GRC)	Social reading comprehension (SRC)	Work reading comprehension (WRC)	Social task completion (ST)	Work task completion (WR)
Variable	V1 Lexicon V2 Syntax V3 Cohesion V4 Gsyntax V5 Glexicon	V6 graph V7 B1 retrieve V8 B1 understand V9 B2 retrieve V10 B2 understand	V14 retrieve V15 understand V16 retrieve V17 understand	V21 retrieve V22 understand V23 retrieve V24 understand	V11 task 1 V12 task2 V13 task 3	V18 task 1 V19 task 2 V20 task 3

b. Confirmatory factor analysis

In order to determine the inter-relationships among scores of the GEPT and TBT and the sub-constructs to be measured, multiple CFA models were proposed, estimated, and evaluated. A covariance matrix of variables was analyzed using the maximum likelihood (ML) estimation method for the structural model described below. In total, six plausible models were tested to examine the relationships among the underlying factor structures across the GEPT and TBT reading tests. These six models were chosen because they were expected to reveal in what ways and to what extent the twenty-four variables related to each other, and in what ways and to what extent they relate to the six subsection factors and the two tests. In other words, these six models revealed the various underlying relationships among the two tests and the 24 variables, and tested to what extent the six subsection factors and the method effect were involved in these relationships. These models are described briefly below and are illustrated in Figures 3.1-3.6.

1. Model: Unitary first-order factor model (UFM): This model hypothesized that a single unitary trait, "reading ability," is measured by the two tests. Twenty-four variables loaded freely on the single trait with no particular structure, and the distinctness of traits corresponding to the two tests and the six subsections was not tested.
2. Model 2: Correlated first-order factor model (CFM): This model hypothesized that the two tests measure two related but different reading abilities. The variables loaded directly to their matching tests. The distinctness of traits corresponding to the two tests was tested, but the distinctness of traits corresponding to the six subsections was not tested.
3. Model 3: Unitary second-order factor model (USM): This model hypothesized that six first-order factors underlie the variables that comprise the subsections of the two tests, and that a common reading ability underlines the six traits measured by the subsections of the two tests. The variables loaded on their matching subsections. The distinctness of traits corresponding to the six subsections was tested, but the distinctness of traits

corresponding to the two tests was not tested.

4. Model 4: Correlated second-order factor model (CSM): This model also hypothesized that six first-order factor underlie the variables that comprise the subsections of the two tests, but hypothesized that two distinct but correlated reading abilities underlie the six traits measured by the subsections of the two tests. The variables loaded on their matching subsections, and the subsections loaded on their matching tests. The distinctness of traits corresponding to the two tests and the six section factors were both tested.
5. Model 5: Bi-factor model (BIM): This model tested the magnitude of a common reading factor in the presence of the three test factors (GEPT, TBT-social, and TBT-work). The variables were specified to load simultaneously on the common trait as well as the three tests. The two TBT tests were different because they were developed based on tasks sampled from two different TLU domains (social and work domain). The variable loadings on the common trait revealed the extent to which the variables share the same construct, and the loadings to the tests revealed the effect of the methods in the presence of the common trait. Factors tested in the other models could not distinguish the method effect due to the test design, whereas the results of this model could either support or refute the hypothesis that factors tested in other models consisted of the targeted traits or were a method effect. The bi-factor model did not include a higher-order factor structure for the common factor due to some identification problems caused by the confounded trait-method factors within the GEPT test. That is, for the bi-factor model to have a second-order factor, first-order factors needed to be identified apart from the method effects. While in this case, traits could not be extracted from the GEPT test because traits and methods were indistinguishable in this situation.

6. Model 6: Third-order factor model (THM): This model hypothesized that a third-order factor, a common reading ability, underlies the two inter-correlated reading abilities in the correlated second-order factor model (CSM).

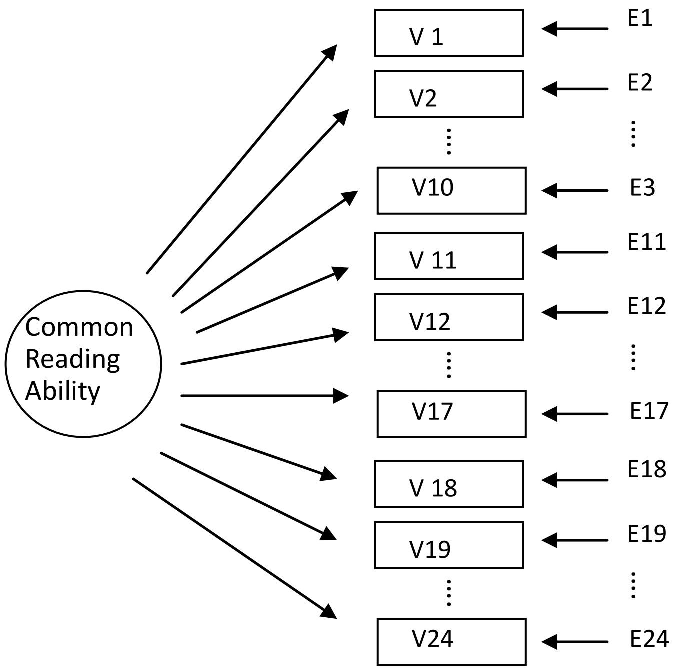


Figure 3.1 Model 1: Unitary first-order factor model (UFM)

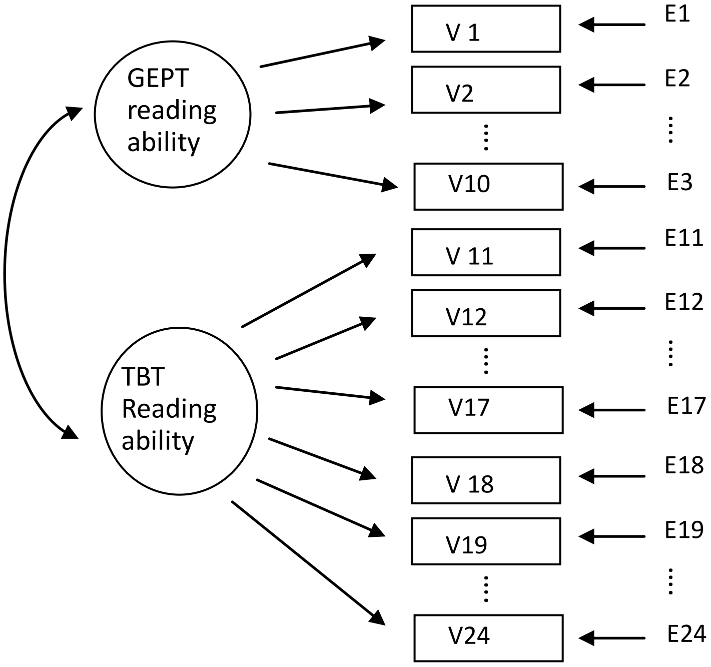


Figure 3.2 Model 2: Correlated first-order factor model (CFM)

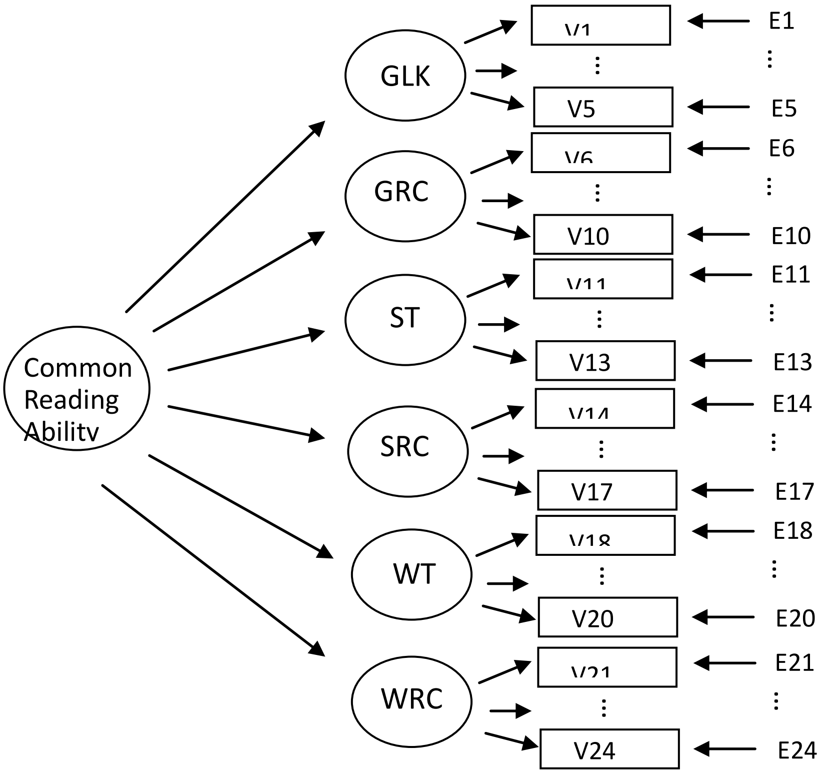


Figure 3.3 Model 3: Unitary second-order factor model (USM)

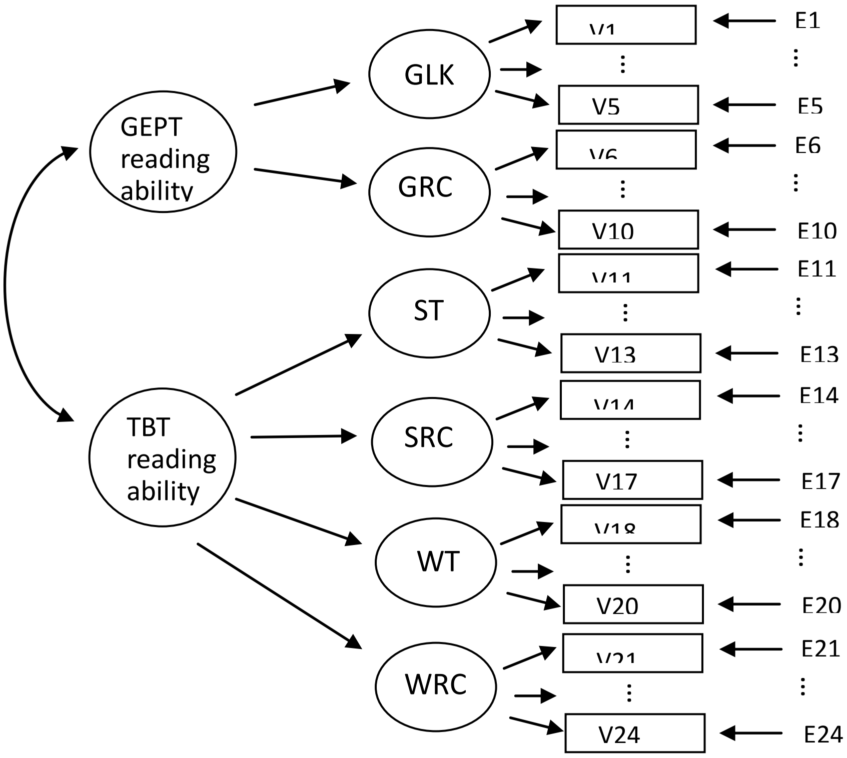


Figure 3.4 Model 4: Correlated second-order factor model (CSM)

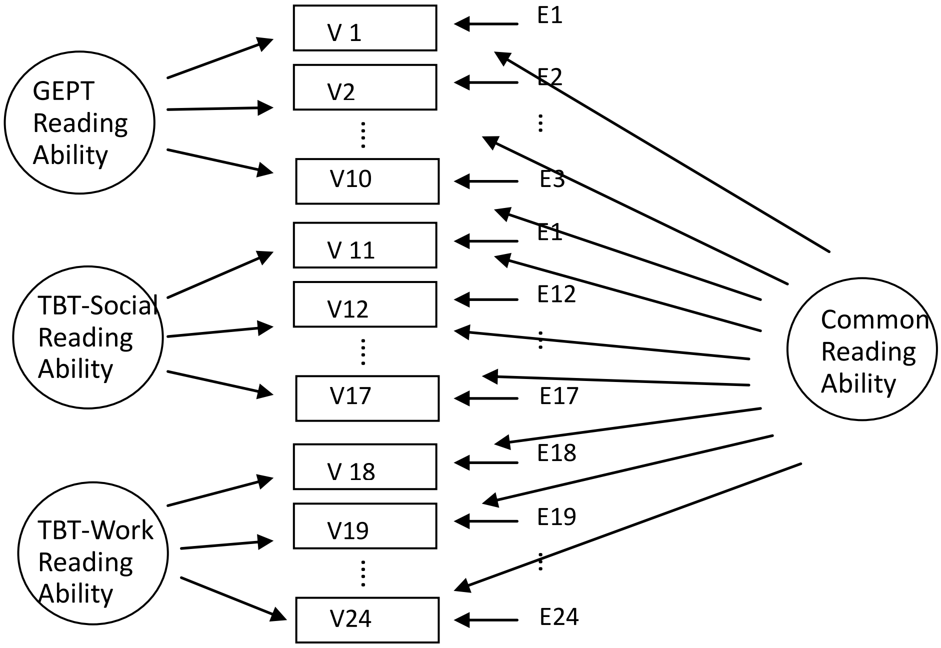


Figure 3.5 Model 5: Bi-factor model (BIM)

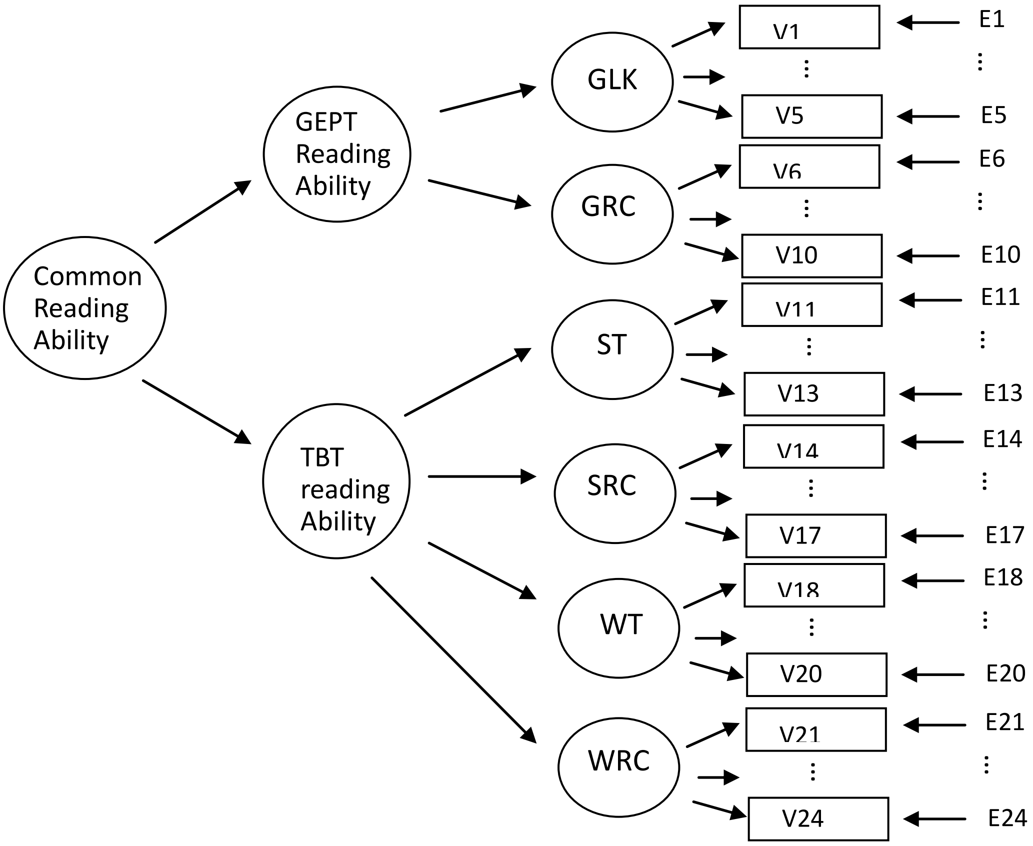


Figure 3.6 Model 6: Third-order factor model (THM)

c. Model selection

The adequacy of the proposed models was evaluated based on multiple criteria: 1) the goodness of the model fit to the data, 2) reasonableness of parameter estimates, and 3) substantive interpretability of the model parameters. The results were evaluated based on the criteria of goodness-of-fit listed below. The chi-square difference test was not employed here because not all the models were nested.

1) *The Satorra-Bentler Scaled Chi-square/df ratio:* This index corrects the effects of a large sample size on the χ^2 statistics. Kline (1998) suggests 2.5 or less as an acceptable model fit

2) *Non-Normal Fit Index (NNFI):* This index avoids extreme underestimation and overestimation. An NNFI of .90 or above indicates an adequate model fit.

3) *The comparative fit index (CFI):* Bentler (1990) proposed this index as a way of avoiding the effects of sample size. A CFI of .90 or higher is considered adequate.

4) *The root mean-square error of approximation (RMSEA):* This index takes into account model complexity as reflected in the degrees of freedom. A RMSEA value less than .05 is considered as an indication of a close fit, and a value from .05 to .08 is an acceptable fit (Browne & Cudeck, 1993).

5) *Standardized Root Mean Squared Residual (SRMR):* A very small SRMR (< 0.05) indicates a strong fit of the model to the data (Hu and Bentler, 1998).

6) *Akaike information criterion (AIC)*: The model with the lowest AIC value is considered to be the best fitting model (Ullman, 2001).

Chapter 4. Results

This chapter describes the results of the study. These will be presented in order of the research questions. First, the results of the qualitative analyses, which addressed the first two research questions, will be described. These include the CEFR level analysis, the text characteristic analysis, and the task/item characteristic analysis. Then, the results of the quantitative analyses will be described. These include factor model evaluation and final model interpretation.

4.1 Results of CEFR level analysis

The first research question investigated the way and extent to which the GEPT reading test construct and content cover the categories and level of descriptors of can-do statements in the CEFR. To address this question, five CEFR categories (reading comprehension, sociolinguistics, text processing, written interaction, and information exchange) and five levels of descriptors (B1, B1+, B2, B2+, C1) were analyzed based on the raters' judgment. Table 4.1 below provides the results of the level analysis. In Table 4.1, the columns represent the texts and items of the GEPT and TBT, while the rows indicate CEFR levels.

Table 4.1: Texts and items of the GEPT and TBT in CEFR levels

	Reading comprehension			
	Texts		Items	
	GEPT	TBT	GEPT	TBT
C2				
C1		Facebook 4.91		
B2+				2
B2	Debussy 4.16 Beverage 4.0	Harvard 4.0 GEOG 3.9 SPUR 3.8	8	23
B1+	Clarkson 3.41		8	3
B1	Imports 3.0		2	
A2+				
A2				
A1				
Average	3.64 B1+	4.16 B2	3.59 B1+	3.95 B2

As can be seen in the table, the ratings and categorizations indicate that raters perceived that the GEPT measured a somewhat lower level of reading proficiency on average (3.64, B1+) than the level it claimed to measure (B2). Half of the texts and items were classified as measuring B2 level of reading comprehension, but the other half of the texts and items were classified as measuring B1 and B1+. For B1 level texts, the “Imports” letter closely reflected the salient features of level B1; it was short, concrete, straightforward, and regularly encountered in a work context. The “Clarkson” article was also straightforward, but it was rated as B1+ because it was slightly longer and was more of a common topic than a regularly encountered topic. For B2 level texts, the “Beverage” and “Debussy” articles were both beyond short, straightforward, and common everyday materials; they reflected some salient features of level B2 such as long, linguistically complex, and reasonably familiar topics. The “Beverage” article reflected another salient feature of level B2—it dealt with a contemporary problem in which the writer adopts a particular stance. The “Debussy” article was rated slightly higher than the “Beverage” article because it was more abstract, and contained a number of low frequency words and figurative speech (imagery and simile).

Table 4.2 presents detailed item level classifications for both the GEPT and TBT, based on the raters' judgments. In the table, the columns represent the CEFR levels and average ratings for each text and item of the GEPT and TBT.

Table 4.2: Item level classifications based on the raters' judgment

GEPT					TBT									
					Social					Work				
Text	CEFR level	Item	CEFR level	Ave. rating	Text	CEFR level	Item	CEF level	Ave. rating	Text	CEF level	Item	CEFR level	Ave. rating
IM	B1	G28	B1	3.16	HV	B2	H1	B2	4.08	GE	B2	G1	B2-	3.83
		G29	B1	3.00			H2a	B2	4.00			G2	B2-	3.83
		G30	B1+	3.33			H2b	B2	4.16			G3	B1+	3.75
CL	B1+	G31	B1+	3.41			H2c	B2-	3.83			G4	B1+	3.66
		G32	B1+	3.58			H3	B2	4.00			GB	B2-	3.83
		G33	B1+	3.41			H4	B2-	3.83	SP	B2	S1	B2-	3.83
		G34	B1+	3.41			HB1a	B2	4.08			S2	B1+	3.66
DE	B2	G35	B2	4.00			HB1b	B2	4.08			S3	B2	3.91
		G36	B2	4.00			HB2a	B2-	3.83			S4	B2	3.91
		G37	B2	4.16			HB2b	B2-	3.83			SB1	B2	4.00
		G38	B1+	3.41	FB	C1	F1	B2	4.16			SB2	B2	4.00
		G39	B2-	3.83			F2	B2	4.00					
BV	B2	G40	B2	3.91			F3	B2+	4.33					
		G41	B2	4.08			F4	B2-	3.83					
		G42	B1+	3.00			F5	B2	4.00					
		G43	B2-	3.75			FB1a	B2	4.16					
		G44	B1+	3.33			FB1b	B2+	4.33					
		G45	B2-	4.00										
Ave	B1+	Ave	B1+	3.59	Ave	B2+	Ave	B2	4.03	Ave	B2	Ave	B2	3.83

*IM=Imports, CL=Clarson, De=Debussy, BV=Beverage, HV=Harvard, FB=Facebook, GE=GEOG, SP=SPUR.

As can be seen from the table, ten GEPT items were rated as level B1 or B1+. Items associated with B1 level texts were all rated as level B1 or B1+ (seven out of the ten), but not all of those associated with B2 level texts were rated as level B2. Three out of eleven items associated with B2 level texts were rated as level B1+ because these items tested more straightforward content than those in the B2 category.

On the other hand, the TBT texts was classified on average (4.16 as in Table 4.1) as measuring the B2 level of reading comprehension, with the majority of the texts (75%) and items (82%) measuring the targeted level. The “Harvard”, “GEOG” and “SPUR” all matched the salient features of B2 level reading, whereas the “Facebook” was classified as level C1 because it was lengthy, complex, embedded with implied attitudes, and contained a number of low frequency words. While one out of the four texts was off level, the items associated with it mostly belonged to level B2 because only the straightforward and essential points in the text were tested, and those relating to C1 level salient features, such as finer points of details and implied attitudes, were not tested.

The proportions of reading points in CEFR levels are presented in Table 4.3. In this table, the columns represent the subsections in the GEPT and TBT, and the rows represent the CEFR levels. The number of points and the percentages were reported in each cell. The second and third columns to the right represent the subtotal points and percentages for the task-based and reading comprehension sections in the TBT. For the GEPT, the numbers of items are indicated, while for the TBT, points are indicated. This is because for the GEPT, each item counts as one point, while some TBT items were worth two points and some tasks were worth four points. It was felt that including points would thus better reflect the proportion for the TBT construct.

Table 4.3: Numbers and percentages of reading points at different CEFR levels

Test	GEPT	TBT						
Section	GRC	SRC	WRC	ST	WT	RC	T	Total
B2+	0	2 (4%)	0	2 (4%)	0	2 (4%)	2 (4%)	4 (8%)
B2	8 (18%)	18 (33%)	9 (17%)	7 (13%)	12 (22%)	27 (50%)	19 (35%)	46 (85%)
B1+	8 (18%)	0	4 (7%)	0	0	4 (7%)	0	4 (7%)
B1	2 (4%)	0	0	0	0	0	0	0
Total	18 (40%)	20 (37%)	13 (24%)	9 (17%)	12 (22%)	33 (61%)	21 (39%)	54 (100%)

*The other 60% (27 pts) of GEPT tested linguistic knowledge (56%) and graph knowledge (4%).

It should be noted that the 18 GEPT items analyzed represent only 40% of the total items in the GEPT (18% at B2, 18% at B1+, 4% at B1 in Table B2). The other 60% of the items were not analyzed because their constructs did not match with the CEFR reading descriptors; 56% of the items measured linguistic knowledge and 4% of the items measured graphic reading. In other words, the GEPT was designed more to measure linguistic knowledge than to measure reading comprehension at the B2 level. This disproportionate construct sampling creates a potentially serious problem for score generalization. With only 40% of the items measuring reading comprehension as defined in the CEFR descriptors and just 18% of these at the intended target level, it is doubtful that the GEPT scores could be interpreted as indicators of the test takers' ability to read at CEFR level B2. For the TBT, for the reading constructs, 61% of the points were rated as measuring reading comprehension and 39% as reading task completion. For levels of reading, the majority (85%) of the TBT points were rated as level B2.

In addition to reading comprehension, these reading items and tasks were also relevant to four other categories of descriptors in the CEFR, namely sociolinguistics competence, text processing, written interaction, and information exchange. These relevant descriptors were included in the analysis because they helped to better portray the very different nature of the constructs involved in the TBT, based on the CEFR. The results of these analyses are presented in Table 4.4, where the number and the proportion of the items involved which CEFR levels of these additional constructs is reported for both tests.

Table 4.4: CEF levels for sociolinguistics, text processing, written interaction, info exchange

Construct		TBT				GEPT			
Subsection		Harvard	FB	GEOG	SPUR	Import	Clarkson	Debussy	Beverage
Socio-Linguistics	comp	3 (50%) 3.33 B1+	0	0	0	0	1 (25%) 3.0 B1	0	1 (17%) 3.0 B1
	task	3 (75%) 3.33 B1+	2 (100%) 4.0 B2	1 (100%) 3.5 B1+	2(100%) 3.5 B1+	--	--	--	--
Text processing	comp	2 (33%) B1	1(20%) B1	2(50%) B1	1(25%) B1	--	--	--	--
	task	2(50%) B1	0	1(100%) B1	0	--	--	--	-
Written interaction	comp	1 (17%) B1	1 (20%) B1	1 (25%) B1	0	--	--	--	--
	task	4 (100%) 3.25 B1	2 (100%) 3.5 B1+	1(100%) 3.0 B1	2(100%) 3.0 B1	--	--	--	--
Info exchange	comp	--	--	--	--	--	--	--	--
	task	4(100%) 3.5 B1+	2(100%) 3.5 B1+	1(100%) 3.5 B1+	2(100%) 3.5 B1+	--	--	--	--

*Percentages represent the amount of items in that particular section.

As can be seen in the table, the TBT tasks involved much more (89%) sociolinguistics content at a higher level (B1+) than the TBT comprehension items (16% at B1+) and the GEPT items (11% at B1). This suggests that the reading tasks engaged a much greater extent of sociolinguistic competence than the comprehension items. For “text processing”, one third of the TBT items required some extent of B1 level paraphrasing and summarizing. The TBT items and tasks required test takers to actively paraphrase specific points in the texts and produce their written responses. The GEPT items were not counted; while the GEPT asked questions (11%) that required paraphrasing, the multiple-choice options given had changed the nature of paraphrasing to mere recognition and understanding. For “written interaction”, the GEPT items did not require any use of language, whereas the TBT required test takers to briefly explain the problems involved in the texts. More specifically, the TBT tasks required the most explanation (100% of the tasks), and the TBT comprehension items required only limited explanations (15% of the items). For “information exchange”, none of the GEPT and TBT comprehension items required test takers to exchange information.

However, the TBT-social tasks required test takers to synthesize and report information and arguments (B2) and give opinions (B1) to demonstrate their understanding of the texts, and the TBT-work tasks required test takers to give a description of how to carry out a procedure (B1) with reasonable precision reliably (B2), and to reliably advise on matters related to occupational roles (B2) and check and confirm factual routines (B1).

Table 4.5 presents a graphic profile of the relationship of the GEPT and TBT to the CEFR levels in five categories.

Table 4.5: Graphic profile of the relationship of GEPT and TBT to CEFR levels

	Reading comprehension		Sociolinguistics		Text processing		Information exchange		Written interaction	
	GEPT	TBT	GEPT	TBT	GEPT	TBT	GEPT	TBT	GEPT	TBT
C2 ↑										
C1										
B2+										
B2		100%								
B1+	40%	▨		39%				32%		
B1	▨	▨	4%	Task		32%		Task		41%
A2+	▨	▨	▨	Mostly		▨		only		Task
A2	▨	▨	▨	▨		▨		▨		Mostly
A1	▨	▨	▨	▨		▨		▨		▨

*Percentages represent the total amount of test items involved.

As can be seen in the table, the abilities involved in the GEPT items were different from those involved in the TBT items and tasks. The GEPT items engaged a considerable extent (40%) of B1+ “reading comprehension” and a very limited extent of B1 “sociolinguistics” (4%). The TBT-social comprehension items (text comprehension and interpretation) engaged a great extent of B2 “reading comprehension”, some extent of B1 “sociolinguistics” and “text processing”, and a limited extent of B1 “written interaction”. The TBT-work comprehension items (in contrast to task completion) engaged a great extent of B2 “reading comprehension”, a moderate extent of B1 “text processing”, and a limited extent of B1 “written interaction”. The TBT tasks thus engaged a great extent of “reading comprehension”, “sociolinguistics”, “information exchange”, and “written interaction”, and some extent of “text processing”. These varying degrees of engagement of these constructs across the two tests reveal that the GEPT items were somewhat different from the TBT comprehension items and were very different from the TBT tasks. In other words, the TBT, especially those task-based items, engaged a wider range and extent of abilities than the GEPT comprehension items.

This section addressed the first research question and investigated the ways and extent to which the GEPT reading test construct and content cover the categories and level of descriptors of can-do statements in the CEFR. It was found that the GEPT reading comprehension texts and items were rated as measuring a lower level of reading proficiency (B1+) than the level it claimed to measure (B2). Also, the GEPT was designed more to measure linguistic knowledge than to measure reading comprehension at the B2 level. The difference between the nature of the TBT items and tasks and those of the GEPT items was found significantly large in terms of the construct of reading, sociolinguistics, text processing, written interaction, and information exchange as specified in the CEFR. It appears that the GEPT items engaged a much narrower range and extent of abilities included in the CEFR than did the TBT.

4.2 Results of text characteristics analysis

The second research question investigated the way and extent to which the characteristics of GEPT reading test tasks corresponded to those of the TBT tasks derived from the can-do statements in the CEFR and the TLU domain. The characteristics of the texts and tasks/items were also analyzed based on the raters' judgment. The results of text characteristics analysis are presented in Table 4.6, where raters' average ratings and classifications for the nine text characteristics variables are reported for each text. The last three columns report the average ratings of the nine characteristics variables for the three tests, TBT-social, TBT-work, and the GEPT. All ratings were on a scale of 1 to 4. The second column to the right indicates the extent of similarities and differences between the average ratings of the TBT and the GEPT.

The ratings and categorizations in Table 4.6 reveal that the texts of the GEPT and TBT are somewhat different for "linguistic complexity", "domain", "discourse type", "abstractness", "rhetorical organization and features", and "proposition density". On the other hand, they are quite different in terms of "proposition complexity" , "pragmatics", "sociocultural specificity", and "amount of input". Among the relative differences, the "linguistic complexity" of the texts in both tests was rated similarly for lexicon (both ranging from 2 to 4 with averages of 2.75 and 3, for GEPT and TBT, respectively) and syntax (ranging from 2 to 3 and 3.5 with averages of 2.88 and 2.75). However, taking domain into consideration, the TBT-social texts were linguistically most complex (3.38), followed by the GEPT texts (2.88), while the least complex were TBT-work texts (2.25). It would appear that the texts from the work domain were linguistically less complex (2 for GEOG and Import, 2.5 for SPUR) than the texts from the personal and educational domains (all the rest of the texts ranges from 3 to 3.75).

Table 4.6: Results of text characteristics rating (based on Dutch Grid, CEFR, and Carr)

Feature/test	TBT-social		TBT-work		GEPT				TBT		vs	GEPT
Passage	Harvard	Facebook	GEOG	SPUR	Import	Clarkson	Debussy	Beverag	Social	work		GEPT
1. Domain	Personal	Personal	Occupation	Occ	Occ	Edu	Edu	Personal	Per	Occ.	≠	Mixed
2. Discourse type (A/E/D/N/I)	A3 (D2,E2)	A3, E3 (D2, N2)	E3 (D2.5, I2.5)	D3, E3 (I2)	E3	N3 (E2)	E3 (A2, D2)	A3, E3	A3, E3	E3, D3	≠	E3, N3, A3
3. Abstractness	2.2	2.2	1.25	1.3	1	1	2.3	1.3	2.2	1.28	>	1.4
4. Linguistics	3	3.75	2	2.5	2	3	3.5	3	3.38	2.25	≠	2.88
Lexicon	3	4	2	2	2	3	4	3	3.5	2	>	3
Syntax	3	3.5	2	3	2	3	3	3	3.25	2.5		2.75
5. Rhetorical												
Organization	2.75	3	3	2	2	2.5	2.5	2.5	2.88	2.5	=	2.37
Features	4 types Example 3 Contrast 2 Problem 2 Analysis 2	4 types Example 2 Constrast 2 Cau/eff2 Analysis3	1 type Problem 3	1 type Classify 3	None	1 type Example 3	4 types Example 3 Contrast 2 Cau/Eff 2 Analysis 2	1 type Example 3	4 types	1 type	>	vary
6. Proposition												
Density	3	3	1	2.75	2	3	2.5	3	3	1.87	=	2.6
Complexity	2	3	3.5	3	1	1.5	2.5	1.5	2.5	3.25	>	1.6
7. Pragmatics												
Directness	0	0	1	4	4	0	0	4	0	2.5	>	1
Speech act #	0	0	3	3	(3*)	0	0	1	0	3	>	.25
Lang. Function	ID	ID	MA/ID	MA	ID/MA	ID	ID	ID	ID	MA	>	ID
8. Sociocultural			College	US								
Specificity	US 2.5	US 3	Enroll 3.5	Colleg 4	None	None	Art 2	None	Highly	Highly	>	Little
Lang. use	Some	A lot	A couple	some	None	None	Some	None	More	Some	>	A few
Formality	formal	For/causal	For/causal	Formal	Formal	Formal	Formal	Formal	Mixed	Mixed	>	Formal
9. Amount of input												
# of words in texts	415	726	619	941	166	252	357	410	1141	1560	>	1185
# of sentences	14	43	31	48	7	15	19	20	58	79	=	61
Ave. sent. length	29.6	16.8	19.9	19.6	23.7	16.8	18.7	20.5	19.6	19.7	=	19.4
# of words in Qs	155	187	139	221	103	130	186	170	342	360	<	587

*Imports too short to be compared with GEOG and SPUR.

For “domain”, both tests categorized texts as occupational (50% for TBT, 25% for GEPT) and personal (50% for TBT, 25% for GEPT), but the GEPT had two texts (50%) that were educational and the TBT had no texts from the educational domain. For “discourse type”, the GEPT texts were mostly expository involving some narration and argument (75%); the TBT texts were also mostly expository (75%), but involved more argument in the TBT-social texts (50%) and more description in the TBT-work texts (25%). For “abstractness”, both the TBT-work and the GEPT texts (except the Debussy article) were rated mainly concrete (ranges from 1 to 1.3), whereas both texts in the TBT-social were found to involve some degree of abstractness (2.2). For “rhetorical organization”, the Facebook and GEOG texts in the TBT were found to have a more complex structure (3), while the GEPT texts shared a similar level of complexity (average of 2.37). For “rhetorical features”, the TBT-social texts contained a wider variety of rhetorical features (4 variations) than the TBT-work and the GEPT texts (1 variation). Only the Debussy article in the GEPT used as many varieties as those in the TBT-social. For “proposition density”, most of the GEPT and TBT texts were rated as moderately dense (all in 3, Debussy 2.5), while the Precision Imports passage was moderately spare (2) and the GEOG email was highly spare (1). It seems that the propositions presented in the business exchanges were more scattered than those in other prose.

Among the major differences, the TBT texts (average of 2.88) were propositionally much more complex than the GEPT texts (average of 1.6). The TBT texts were rated as moderately complex and the GEPT texts were relatively straightforward. In fact, one of the key text features for B2 reading in the CEFR was “propositionally and linguistically complex”; it seems that the GEPT is linguistically comparable with the TBT, but it was definitely not as propositionally complex as the TBT. For “pragmatics”, the TBT-work texts (GEOG and SPUR) involved a larger number of speech acts and manipulative functions than the GEPT texts (Precision Imports and Functional Beverages). These texts contained mostly direct speech acts and a few indirect ones except for the GEOG email exchange, which had quite the contrary feature. It involved the use of more indirect (75%) than direct speech acts (25%). These indirect requests in the GEOG were used by the interlocutors to politely make requests in an occupational setting. Questions and concerns in this setting were functioned as requests and directives; if one were to mistake its manipulative function for an ideational function, he/she might misread and therefore fail the TBT work task.

For “sociolinguistics”, the TBT texts involved more cultural references, figures of speech, and register shift than the GEPT texts. In the TBT-social, the texts were US culture specific and figures of speech were embedded. For instance, the Harvard article discussed the concepts of the American dream, selective universities, and college admission policy that were all US culture specific. The Facebook article mentioned US figures such as Julian Assange and the Tea Party and it additionally used a number of metaphors and irony. The texts in the TBT-work, on the other hand, were US college culture specific. For instance, the GEOG text discussed enrollment procedures and included the use of college culture specific acronyms like BruinBill and GIS&T institute. The SPUR text described application requirements and procedures, and included the use of US college culture references such as tracking, historically underrepresented groups, US permanent residents, and cumulative GPA. In contrast, the GEPT texts were all general and not specific except for the Debussy article that was moderately art specific and used similes. For “amount of input”, the GEPT texts (average of 296 word/passage) were much shorter than the TBT texts (average of 675 word/passage). Notably, the GEPT multiple-choice questions (stems and options) contained 683 words, which was more than any of the GEPT reading passages.

In summary, compared to the GEPT texts, the TBT-social texts were more abstract and argumentative, with more complex and varied rhetorical organizations and features, linguistically more complex, not educational, and sociolinguistically more complex, and the TBT-work texts were more pragmatically and sociolinguistically complex. However, it was not these characteristics or the linguistic complexity that set the TBT and the GEPT texts apart. Rather, the amount of complex propositional content that needed to be understood and the pragmatic and sociolinguistic specificities associated with the texts made them quite different from each other. In other words, the TBT texts were all much longer and propositionally and sociolinguistically more complex than the GEPT texts.

4.3 Results of item characteristics analysis

The results of item characteristics analysis are presented in Table 4.7, where raters' average ratings and classifications for the ten item characteristics variables are reported for each text. The last three columns report the average ratings of the ten characteristics variables for the TBT and the GEPT. The averages for the TBT comprehension items and task items are reported separately. All ratings were on a scale of 1 to 4. The second column to the right indicates the extent of similarities and differences between the average ratings of the TBT and the GEPT.

The ratings and categorizations in Table 4.7 reveal that the items of GEPT and TBT share a similar level of "input linguistic complexity" (both ranging from 2 to 3.5 with

Table 4.7: Results of item characteristics rating (based on Dutch Grid, CEFR, and Bachman)

Feature/test	TBT-social				TBT-work				GEPT				TBT overall		vs	GEPT
Passage Comp vs. task	Harvard SRC	ST	Facebook SRC	ST	GEOG WRC	WT	SPUR SRC	WT	Impo rt	Clark son	Debu ssy	Beverag e	TBT RC	T		GEPT
1. Input linguistics Lexicon Syntax	2.6 2.7 2.5	3 3 3	3.4 3.8 3	3 3.5 2.5	2 2 2	2 2 2	2.1 2 2.2	2.5 2 3	2 2 2	3 3 3	3.5 4 3	3 3 3	2.5 2.6 2.4	2.6 2.6 2.6	=	2.87 3 2.75
2. Content tested Main idea/ Detail/Infer	D1 IM2 ID3	M2 D1 IM1	D5 ID1	D1 ID1	M2 D2	D1	D3 IM1	D2	M1 D1 ID1	M1 D2 ID1	D3 ID2	M1 D2 IM2/ID1	M2 D11 I7	M2 D5 I2	=	M3 D8 I 7
3. Implicitness	2.2	2.5	1.6	2.9	1.1	1.5	1.3	1.8	1.5	1.8	2	1.8	1.6	2.2	= >	1.8
4. Abstractness	2	2.75	2.4	2.5	1.1	1.5	1.5	1.5	1.2	1.75	2.3	1.7	1.75	2.1	= >	1.74
5. Operation recog/interpret /analyze	2.1	2.4	1.5	2.65	2	2.5	1.4	2.5	1.7	1.5	1.7	1.7	1.75	2.5	= >	1.65
6. Amount of input	2.8	3.5	2.16	2.25	1.87	2	2.2	4	(3*)	2.1	1.7	2.3	2.27	2.94	= >	2
7. Lang. function Input Response	 ID ID	 ID ID	 ID ID	 ID ID	 M/ID ID	 M M	 M ID	 M M	 ID None	 ID None	 ID None	 ID None	 ID/M ID	 ID/M ID/M	 > >	 ID None
8. Sociolinguistics % of items CEFR rating CEFR level	 50% 3.33 B1+	 75% 3.3 B1+	 0 0 0	 100% 4.0 B2	 0 0 0	 100% 3.5 B1+	 0 0 0	 100% 3.5 B1+	 0 0 0	 25% 3.0 B1	 0 0 0	 17% 3.0 B1	 16% 3.3 B1+	 89% 3.6 B1+	 = >	 11% 3.0 B1
9. Strategic demands	1.5	2.3	1.2	2.5	1.5	2.7	1.3	3.5	0	0	0	0	1.38	2.75	>	0
10 Response linguistics Lexicon Syntax	 1.25 1.5 1	 1.5 2 1	 0.9 5 0.9 1	 1 1 1	 1.4 1.4 1.4	 1.25 1.5 1	 1.2 1 1	 1.25 1.5 1	 0 0 0	 0 0 0	 0 0 0	 0 0 0	 1.2 1.26 1	 1.25 1.5 1	 >	 0 0 0

*"Import" text is too short to be taken into account.

averages of 2.87 and 2.55, for GEPT and TBT, respectively) and the type of “content tested”, and they are to some extent comparable according to “implicitness” and “abstractness”; however, they are quite different in terms of “operation”, “amount of information” , “language functions”, “sociolinguistics”, and “strategic demands”. Among the similarities, “linguistic complexity” for both test items was rated similarly for lexicon (ranges from 2 to 4/3.8 with averages of 3/2.6) and syntax (ranges from 2 to 3 with averages of 2.75/2.5). But taking domain into consideration, TBT-work was linguistically less complex than TBT-social and the GEPT, and it was in fact linguistically aligned with the B1 level passage “Precision Import” in the GEPT. For “content tested”, both the GEPT and TBT have similar percentages of items measuring the content of main ideas (14% and 17% for GEPT and TBT, respectively), details, and inferences. For “implicitness” and “abstractness”, the GEPT and TBT-work items were rated mostly explicit and concrete, whereas TBT-social items were found to involve some extent of implicitness and abstractness, mostly due to the task-based rather than the comprehension items.

Among the major differences, the TBT items were rated utilizing a wider range of “operations” (ranges from 1.4 to 2.65) than the GEPT items (ranges from 1.5 to 1.7). That is, the TBT items involved more interpretation and analysis (average of 2.1), whereas the GEPT items involved mostly recognition (average of 1.65). Only the Facebook and SPUR comprehension items in the TBT were similar to the GEPT items as they both involved mostly recognition, but the remaining items in the TBT required mostly interpretation and analysis. More specifically, more items of TBT (57%) involved interpretation and analysis and fewer items (43%) involved recognition, whereas more items of GEPT (61%) involved recognition and fewer items (39%) involved interpretation and analysis. Notably, the TBT task-based items involved a larger amount of interpretation and analysis (89% of items rated above 2 with an average of 2.5) than the TBT comprehension items (42% of items rated above 2 with an average of 1.75). In other words, the differences between the GEPT and TBT were mainly due to those task-based items rather than comprehension items.

For "amount of information", the TBT required more information to be processed (average of 2.6 for beyond a paragraph) than the GEPT (average of 2 for within a paragraph). That is, the TBT had more items (45%) testing content beyond a paragraph and fewer items (20%) testing localized information, while oppositely, the GEPT had more items (47%) testing localized information and fewer items (27%) testing content beyond a paragraph. It is noteworthy that 67% of TBT task-based items required test takers to process almost the entire passage. For "language functions", the TBT items were categorized as having a wider range of functions than the GEPT items. TBT-work items involved both ideational and manipulative functions for the input and response, whereas the GEPT items involved only the ideational function for the input and no functions at all for the response because there was no language function to be performed in answering multiple-choice questions.

For "sociolinguistics", the TBT items contained a larger proportion (89%) and a higher level (B1+) of sociocultural content than the GEPT items (11% at B1). This difference is mainly due to the task-based items; many of these items required test takers not only to process texts that were sociocultural specific and contextualized but also to perform and act appropriately according to those specificities and contexts. For instance, they had to understand the implication of American dream with equality to be able to respond to the prompt, and they had to identify the relevance of an individual's comment on the American dream to the text and provide his/her own opinion; also, they had to sort out a clerical procedure of departmental scholarship application based on three office email exchanges and report back to the supervisor. These types of tasks were quite different from those of the typical comprehension items for they required test takers to take culture, context, interlocutors, and appropriateness into consideration to demonstrate their understanding of the texts.

For “strategic demands”, the TBT items required some extent of strategic assessment and planning in order to produce the responses, whereas strategic assessment was not required by the GEPT items to produce the responses because of the multiple-choice test formats. The fill-in-blanks questions and short answer questions in the TBT comprehension required limited strategic efforts (1.38), while the TBT situational task items required a greater extent of assessment and planning (2.75) to complete. For “response linguistic complexity”, the TBT items required test takers to produce responses by supplementing frequent vocabulary and simple structures in addition to the content in the texts, whereas the GEPT items did not require any use of language. It is noteworthy that while TBT items involved writing, especially those task items, the level of linguistic complexity required was fairly low. The scoring criteria were content, appropriateness, and comprehensibility. Linguistic errors not hindering comprehension of the response content were acceptable.

In summary, compared to the GEPT items, the TBT-social items were somewhat more abstract and implicit, and the TBT-work items were more functionally complex. However, it appears not to be the linguistic complexity or the topical contents that distinguishes the GEPT and TBT items. Rather, it is the amount of input needed to be interpreted rather than recognized and the pragmatic and strategic demands associated with the text and tasks that make them quite different from each other. In other words, the TBT involves a wider range and extent of characteristics, competences, and input than the GEPT, and the task-based items, in particular, are the ones that contribute to this difference.

The qualitative results of the study show that the nature and level of the GEPT constructs and the characteristics of the GEPT texts and items were both quite different from those of the TBT. The GEPT reading comprehension texts and items were rated at B1+ level, whereas the TBT texts and items were rated at B2 level. Regarding constructs of sociolinguistics, text processing, written interaction, and information exchange, the GEPT items engaged very little of these abilities, while the TBT, especially those task-based items, engaged a great extent and variety of these abilities. Also, the GEPT texts were rated much shorter and propositional and pragmatically less complex than the TBT texts, and the GEPT items engaged a narrower range and extent of characteristics, competences, and input to be processed than the TBT item and tasks.

4.4 Results of confirmatory factor analysis

The analyses in this section address the third research question and investigated in what ways and to what extent the underlying construct of the GEPT reading test corresponds to that of the TBT developed based on the can-do statements in the CEFR and the TLU domain. The underlying factor structures of both tests were inspected using confirmatory factor analysis. The six models that were proposed in Chapter 3, section 3.6.2 were empirically tested to examine the relationships among the underlying factor structures across the GEPT and TBT reading tests. The six models were chosen because they were expected to reveal in what ways and to what extent the twenty-four variables related to each other, and in what ways and to what extent they relate to the six subsection factors and the two tests.

4.4.1 Model comparisons

Five pairs of models were compared to reveal the correspondences of the factor structures between the GEPT and TBT. The first comparison was between the unitary first-order factor model (Model 1, UFM) and correlated first-order factor model (Model 2, CFM). This comparison was intended to reveal the distinctness between the GEPT and TBT when the six subsection factors were not present. Next, the unitary second-order factor (Model 3, USM) model was compared with correlated second-order factor model (Model 4, CSM) in order to reveal the distinctness between the GEPT and TBT when the six subsection factors were present. Correlated first-order factor model (Model 2, CFM) was then compared with correlated second-order factor model (Model 4, CSM) to reveal whether the six subsection factors were present. This was followed by a comparison between the correlated second-order factor model (Model 4, CSM) and bi-factor model (Model 5, BIM) to reveal the magnitude of method effect in the presence of a general reading factor. Finally, the correlated second-order factor model (Model 4, CSM) was compared with the final model, the third-order factor model (Model 6, THM), to reveal whether there was a common reading factor underlying the two correlated factors. The results of these comparisons are presented in Table 4.8. Recall from Chapter 3 that the chi-square difference test was not used because not all of the models were nested.

Table 4.8: Summary of CFA model testing

Proposed models		Model df	SB Scaled χ^2	SB χ^2 /df	NNFI	CFI	RMSEA	SRMR	AIC
1.	Model 1: UFM	252	456	1.8	.89	.90	.058	.05	-47
	Model 2: CFM	251	406	1.6	.92	.92	.047	.051	-95
2.	Model 3: USM	246	378	1.53	.93	.938	.052	.054	-113
3.	Model 4: CSM	245	359	1.46	.94	.946	.05	.044	-130
4.	Model 5: BIM	228	327	1.43	.94	.95	.042	.043	-128
5.	Model 6: THM	245	360	1.46	.939	.946	.051	.044	-129

1. Testing for distinctness between the GEPT and TBT reading ability in first-order factor models: Model 1 (UFM) vs. Model 2 (CFM)

This comparison was intended to reveal the distinctness of reading abilities tested in the GEPT and TBT reading tests when the six subsection factors were not present. As listed in Table 4.8, the values of SB χ^2/df, NNFI, CFI, RMSEA, and SRMR for both models show that Model 2 (CFM) fits better than Model 1 (UFM). The AIC index also indicates that Model 1 better represents the relationship between the GEPT and TBT. Accordingly, the distinctness between the GEPT and TBT reading abilities is affirmative when the six subsection factors are not included. In other words, the variables load in a particular structure to their matching tests; they measure different underlying constructs rather than a single trait.

2. Testing for distinctness between the GEPT and TBT reading ability in second-order factor models: Model 3 (USM) vs. Model 4 (CSM)

This comparison tested the distinctness of reading abilities measured in the GEPT and TBT reading tests when the six subsection factors were present. As listed in Table 4.8, the values of SB χ^2/df, NNFI, CFI, RMSEA, and SRMR for both models show that these models fit the data well. While the difference between the two models remains small, Model 4 (CSM) fits a little better than Model 3 (USM). The AIC index also indicates that the difference exists. Thus, the distinctness between the GEPT and TBT reading abilities was found to be affirmative when the six subsection factors were present. In other words, the six subsection factors measured different underlying constructs rather than one common trait.

3. Testing for the presence of the six subsection factors: Model 2 (CFM) vs. Model 4 (CSM)

Since the reading abilities measured in the GEPT and TBT were found to be distinct from each other whether the six subsection factors were present or not, this comparison was intended to examine which model, Model 2 (CFM) or Model 4 (CSM), better represents the relationships of the underlying constructs between the two tests. As listed in Table 4.8, the values of SB χ^2/df, NNFI, CFI, RMSEA, and SRMR for both models show that Model 4 (CSM) fits slightly better than Model 2 (CFM). The AIC index also indicates that Model 4 better represents the relationship between the GEPT and TBT. Hence, the two reading abilities were found to be distinct when the six subsection factors were present rather than when they were absent. Model 4 closely reflects the test specifications of the two tests in that the reading ability measured in the GEPT consists of linguistic knowledge and reading comprehension, and that ability is related to but different from the reading ability measured in the TBT, which consists of social reading comprehension, social task completion, work reading comprehension, and work task completion.

4. Testing for the magnitude of the method effect in the presence of a general reading factor: Model 4 (CSM) vs. Model 5 (BIM)

Model 4 (CSM) hypothesized that the covariance among the 24 observed variable can be explained solely in terms of trait factors—six primary traits and two higher-order traits. However, a test method effect is a possible source of construct irrelevant variance. In order to claim that the trait factors proposed in Model 4 (CSM) alone account for the covariance among the 24 variables, it is necessary to rule out the possibility of a significant method effect. This can be done by specifying a model that includes factors for test methods as well as abilities, i.e., Model 5 and comparing this with Model 4.

The fit indices of SB χ^2/df, NNFI, CFI, RMSEA, and SRMR for both models in Table 4.8 show that these models fit the data well. While the difference between the models remains minimal, the bi-factor model fits slightly better than Model 4 (CSM). An examination of the parameter estimates in the bi-factor model also shows that a majority of the variable loadings on the general reading factor is comparatively higher than those on the method factors. However, although the general reading factor has been successfully extracted, the variances left to be explained are only those related to the GEPT method effect. Half, five out of the ten, variable loadings on the GEPT test method factor are significant (V2 syntax, V3 cohesion, V4 gap-fill syntax, V7-B1 retrieve, and V10-B2 understand), whereas none of the rest of the variable loadings on the TBT-social and TBT-work test method factors is significant. The method factors could stand alone due to the method effect of the GEPT, but not those of the other tests. The relatively larger loadings of these variables (.25–.44) of the GEPT indicate that they measure something else, in addition to or other than the general reading ability specified in the construct.

While the bi-factor model (Model 5, BIM) revealed the magnitude of the method effect in the GEPT, the interpretability of this model is not as meaningful as that of the correlated second-order factor model. In the correlated second-order factor model, all the variables load significantly onto their corresponding subsection factors. In addition, the correlated second-order factor provides more substantive support to the underlying trait construct than the bi-factor model because it accounts for the relationships among the trait factors, whereas the bi-factor model does not capture such relationships. Therefore, the correlated second-order factor model was a preferable model representing the relationship between the GEPT and TBT.

5. Testing for the presence of a common reading factor underlying the two correlated factors: Model 4 (CSM) vs. Model 6 (THM)

Model 4 (CSM) was compared with Model 6 (THM) to test for the presence of a common factor underlying the two correlated factors. Because both models yield the same number of parameters to be freely estimated, the two-factor model and its corresponding third-order model are statistically indistinguishable in this case. However, the third-order factor model is preferred to the correlated two-factor model because it is more interpretable. That is, the third-order factor model makes it easier and more direct and explicit to interpret the relationships between the two factors. It better accounts for the relationships between the two correlated factors by imposing a common factor that explains the complex inter-relationships between them, whereas the correlated two-factor model leaves the relationship unexplained.

Based on the analyses of model comparisons, it was found that neither the twenty-four variables nor the six subsection factors that comprised the two tests loaded on one single trait. On the contrary, these variables and subsections loaded on their respective distinct subsections and tests, along with a common reading ability that underlay both tests. Model 6 (THM) best represents the correspondence of the underlying construct between the GEPT and TBT.

4.4.2 Interpretation of the Third-Order Factor model (THM)

Since Model (THM) provided a good fit to the data and was the most interpretable model, the specific parameter estimates of this model need to be interpreted. Table 4.9 presents the standardized loadings of the six first-order factors (GLK, GRC, ST, SRC, WT, WRC) on the two second-order factors (GEPT reading and TBT reading) and the loadings of the two second-order factors on the third-order factor (common reading) in Model 6 (THM).

Table 4.9: Standardized loadings of first-order factors on the second & third-order factors

Factors	GEPT reading	TBT reading	Common reading	Error	SMR
GEPT Linguistic knowledge (GLK)	.94			.32	.89
GEPT Reading comprehension (GRC)	1.00			.00	1.00
TBT Social task completion (ST)		.93		.34	.88
TBT Social reading comprehension (SRC)		1.00		.00	1.00
TBT Work task completion (WT)		.83		.55	.68
TBT Work reading comprehension (WRC)		.91		.41	.83
GEPT reading			.91	.40	.83
TBT reading			1.00	.00	1.00

SMR= Squared multiple correlation (R-squared)

As can be seen in Table 4.9, both tests, GEPT and TBT, had high loadings, .91 and 1.00, respectively, on the third-order factor “Common Reading Ability”. This indicates the presence of a common reading ability underlying both tests. It is notable that the loading of “common reading” to “TBT reading” and the loading of “TBT reading” to “TBT social reading comprehension” are both 1.00. This indicates that both “common reading” and “TBT reading” represent the same trait as measured in “TBT social reading comprehension.” While other test and subsection factors measure similar common reading abilities, they are not identical as these factors are. Among the subsection factors, the loadings of “GEPT linguistic knowledge” (.85) and “TBT work task completion” (.83) are lower than the loadings of other factors to “common reading” (.91 to .93). While this difference might seem small, it is not trivial; it indicates that these two factors measure some other abilities that differ from the targeted competence—social reading comprehension.

This finding corresponds to the previous model comparisons, indicating that the GEPT and TBT measure similar yet distinct reading abilities and that the six subsection factors have their own varied relationships to the common underlying ability.

To further inspect the associations among the six subsection factors, a correlation matrix was obtained by multiplying a combination of these higher-order factor loading that lie between them. The correlation matrix in Table 4.10 shows the relationships between pairs of subsection factors.

Table 4.10: Model reproduced correlation matrix of the six subsection factors

	GLK	GRC	SRC	WRC	ST	WT
GLK		.94	.86	.78	.81	.71
GRC	.94		.91	.83	.85	.76
SRC	.86	.91		.91	.93	.83
WRC	.78	.83	.91		.85	.75
ST	.81	.85	.93	.85		.77
WT	.71	.76	.83	.75	.77	

As can be seen in Table 4.10, the associations among the six subsection factors were demonstrated in the model reproduced correlation matrix. The matrix explicates the relationships among these different subconstructs in reading.

The correlations show that "linguistic knowledge" is closely related to "GEPT reading comprehension" (.94) and a little less to "TBT social reading comprehension" (.86), but the correlations drop substantively with "work reading comprehension" (.78), "social task completion" (.81), and "work task completion" (.71). This reveals that linguistic knowledge is mostly relevant to general text reading but not to the contextualized reading texts and situational types of reading tasks. In other words, when the reading test items become more task-based and work/context specific, the less association they have with discrete linguistic knowledge per se. This also suggests that "work reading comprehension," "work task completion," and "social task completion" involve some specific language abilities in addition to general reading comprehension. Furthermore, the correspondence between "linguistic knowledge" and "work task completion" is the weakest among all the correlations (.71, the lowest in the matrix). This supports the existence of a divergence between the two distinct approaches, trait-focused and task-focused, to the conceptualizing construct. Even though they both target B2-level reading and share a common underlying construct, they are still quite different in nature (SMR .50).

On the other hand, the correlation between “GEPT reading comprehension” and “linguistic knowledge” (.94) is higher than the correlation of “GEPT reading comprehension” to all other subconstructs; this might partly due to the method effect, as they are both multiple choice questions, but it can also present a problem in test design because GEPT reading comprehension is supposed to measure reading comprehension more than discrete linguistic knowledge; however, the data shows quite the contrary. Moreover, the loadings also show that “GEPT reading comprehension” correlates the lowest with “work task completion” (.76), indicating that GEPT reading comprehension may measure sufficient social reading comprehension but definitely not serve as a sufficient measure of work tasks that requires test takers to make use of the contextualized texts to solve situational work-related problems.

Then, one would assume that “social task completion” and “work task completion” are strongly correlated since they both are task-based, yet they are in fact rather different (.77). This may indicate that domain/context does have an effect on the underlying construct being measured. That is, tasks can vary to a great extent across domains/contexts. However, that is not the case between “work reading comprehension” and “work task completion.” Even though “work reading comprehension” and “work task completion” are actually based on the same texts, “work reading comprehension” still correlates the highest with “social reading comprehension” (.91) and the lowest with “work task completion” (.75). This indicates that the task type, comprehension questions versus situational tasks, may have a greater effect on the nature of the construct than the text type. This also reveals that “work task completion” may measure some unique traits not measured by other subconstructs because it correlates the lowest to all the other factors. After all, understanding what the text is about does not necessarily mean that one can make use of the text information to solve situational problems in a specific work context. That is, language use and problem solving require some different abilities in addition to reading comprehension, and this is especially the case in a work setting.

The standardized factor loadings of variables on the subsection factors for the third-order factor model are presented in Table 4.11. These factor loadings indicate the strength of the associations between the twenty-four variables and their corresponding subsection factors.

Table 4.11: Standardized factor loadings of variables on the subsection factors in the final model

Test	Subsection	Variable	Name	Reading trait	Error	SMR
GEPT	GLK	1	Lexicon	.49	.87	.24
		2	Syntax	.58	.81	.34
		3	Cohesion	.64	.76	.41
		4	GLex	.73	.67	.54
		5	Gsyn	.57	.81	.33
	GRC	6	Graph	.27	.96	.07
		7	B1 Detail	.68	.73	.46
		8	B1 Und	.44	.89	.19
		9	B2 Detail	.76	.64	.58
		10	B2 Und	.77	.63	.59
TBT- Social	ST	11	Task 1	.56	.82	.31
		12	Task 2	.51	.85	.26
		13	Task 3	.41	.90	.17
	SRC	14	H-Detail	.72	.64	.51
		15	H-Und	.63	.77	.40
		16	F-Detail	.68	.73	.46
		17	F-Und	.61	.78	.37
TBT- Work	WT	18	Task 1	.31	.95	.09
		19	Task 2	.83	.54	.70
		20	Task 3	.79	.61	.62
	WRC	21	G-Detail	.59	.80	.34
		22	G-Und	.75	.65	.56
		23	S-Detail	.74	.67	.54
		24	S-Und	.69	.72	.47

GLK=GEPT linguistic knowledge, GRC= GEPT reading comprehension, ST=TBT social task completion, SRC= TBT social reading comprehension, WT= TBT work task completion, WRC= TBT work reading comprehension. Varibable coding: GLex=gap-fill lexicon, Gsyntax=gap-fill syntax, B1 Und= B1 level understanding, Task= task completion, H= Harvard text, F=Facebook text, G=GEOG email text, S=SPUR text; SMR= Squared multiple correlation (R-squared)

The loadings of all the SRC variables on SRC factors and the loadings of WRC variables on WRC factors are substantial (larger than .50), while the loadings of the GLK, GRC, ST, and WT variables on their related factors yield mix results (three GEPT variables and two TBT task variables are less than .50). This difference indicates that the GEPT variables “lexicon,” “graph,” and “B1 understand” measure some irrelevant constructs and that task items “ST3” and “WT1” share less in common with other task items due to variations in context and situations. The “lexicon” variable which measures the sentence-level completion of a single word has a lower loading (.49), and this indicates that it is the most discrete variable of all linguistic variables. The loading of the “B1 understand” variable is also relatively lower (.44), which may be due to the fact that it targets at a level of difficulty that is not well aligned with the target ability. The “graph” variable in the GRC has the lowest loading of all the loadings (.27), indicating that these graphic items measure very little of the target reading comprehension. For task variables, “ST3” shares less in common with “ST1” and “ST2” because it is based on the Facebook text, while “ST1” and “ST2” are both based on the Harvard text. The same is the case with “WT1,” which measures test takers’ ability to extract specific points from a GEOG email to respond to a supervisor, whereas both “WT2” and “WT3” measure test takers’ ability to draft emails in response to students based on the SPUR text.

One final observation concerns the variables in the linguistic knowledge subsection. It seems that “lexicon” measures more of the target reading trait when it is contextualized in the gap-fill section (.73) than it is in the sentence completion section (.49). This suggests that it is better to test the lexicon in context rather than in isolation. However, this is not the case with syntax; it seems that syntax can be tested regardless of the degree of contextualization (.58 in the sentence completion section and .57 in the gap-fill section).

The results of the final model interpretation indicate that the GEPT and TBT measure similar yet distinct reading abilities and that the twenty-four variables and the six subsection factors have their own varied relationships to the common underlying reading ability. The most distinct subsection factors were the GLK and WT; they had the lowest factor loadings on the common trait, reading comprehension, and the lowest correlation with each other. These loadings reveal the discrepancies that existed among the three distinct constructs of reading, i.e., linguistic knowledge (GLK), reading comprehension (GRC, SRC, WRC), and situated reading tasks (ST, WT). They also show the divergence between the two distinct approaches, trait-focused and task-focused, to the conceptualizing construct.

In summary, the analyses of CFA results show that the twenty-four variables and the six subsection factors load on their respective distinct subsections and tests, along with a common reading ability that underlay both tests. Model 6 (THM) best represents the correspondence of the underlying construct between the GEPT and TBT. The factor correlations suggest that when the items are more task-based, contextualized, and workplace specific, their association with the GEPT tests decreases, especially the discrete linguistic knowledge.

4.5 Chapter summary

This chapter has described and presented the results of the study in order of the research questions. Both the qualitative and quantitative analyses yielded similar results. These suggest that while both the GEPT and TBT tests measure reading comprehension, they are quite different in nature. Not only was the TBT more difficult, the constructs engaged are also more complex. The ratings and classifications of the CEFR levels, additional constructs, and text/task characteristics all correspond to the variations identified in the quantitative analyses. In other words, the findings do not support sufficient correspondences existed between the GEPT and TBT in the aspects of CEFR content/levels, task/text characteristics, and underlying constructs.

Chapter 5. Conclusions

5.1 Conclusions and discussion of the research questions

The first research question investigated in what way and to what extent the GEPT reading test construct and content cover the categories and level of descriptors of can-do statements in the CEFR. To address this question, five CEFR categories (reading comprehension, sociolinguistics, text processing, written interaction, and information exchange) and five levels of descriptors (B1, B1+, B2, B2+, C1) were analyzed based on the judgment of expert raters. For "reading comprehension," the ratings suggest that the GEPT measured a somewhat lower level of reading proficiency (B1+) than the level it claimed to measure (B2). Half of the GEPT texts and items were classified as measuring the B2 level of reading comprehension, while the other half of the texts and items were classified as measuring the B1 and B1+ levels of reading comprehension. The TBT, on the other hand, was judged to measure at the B2 level of reading comprehension, with the majority of the texts (75%) and items (82%) measuring the targeted level. It is noteworthy that the GEPT was designed to measure linguistic knowledge (56% of the items) more than reading comprehension (40% of the items) in order to represent the B2 level reading construct. This disproportionate representation of the TLU construct poses a serious problem for score generalization; with only 18% of total items measuring B2 level reading comprehension, it is questionable that the GEPT scores could be interpreted as indicators of the test takers' ability to read at the B2 level.

The ratings for the additional four CEFR categories reveal significant variation between the nature of the TBT items and tasks and those of the GEPT items. Table 5.1 summarizes the different extent, range, and level of CEFR constructs that were involved in each reading subsection in the GEPT and TBT. Table 5.1 presents the extent to which the six reading subsections involved the construct of reading, sociolinguistics, text processing, written interaction, and information exchange as specified in the CEFR. In Table 5.1, the columns represent the five construct categories in the CEFR that are relevant to the GEPT and TBT, while the rows indicate the six reading subsections measured in the GEPT and TBT. The percentages refer to the proportion of the total items in a subsection that involves a particular construct based on the raters' judgment.

Table 5.1: CEFR constructs across the six subsections in the GEPT and TBT

Test	Sub-section	Reading comprehension	Socio-linguistics	Text processing	Written interaction	Info exchange
GEPT	GLK	0	0	0	0	0
	GRC	**40% B1+**	4% B1	0	0	0
TBT	SRC	**100% B2**	**25% B1+**	**25% B1**	17%	0
	WRC	**100% B2-**	0	**38% B1**	12%	0
	ST	**100% B2**	**83% B1+**	**33% B1**	**100% B1**	**100% B1+**
	WT	**100% B2**	**100% B1+**	**33% B1**	**100% B1**	**100% B1+**

*Percentages represent the amount of items in that particular section.*GLK=GEPT linguistics knowledge, GRC=GEPT reading comprehension, SRC=TBT social reading comprehension, WRC=TBT work reading comprehension, ST=TBT social task completion, WT=TBT work task completion.

As can be seen in Table 5.1, the ratings suggest that the GLK section of the GEPT engaged none of these CEFR constructs; GRC engaged reading comprehension; SRC and WRC involved reading comprehension and text processing. On the other hand, both sections of the TBT—ST and WT—involved all five constructs to a large degree. These varying degrees of engagement of the CEFR constructs across the two tests reveal that the GEPT items were relatively different from the TBT comprehension items and very different from the TBT tasks. In other words, the GEPT items engaged a narrower range and extent of abilities included in the CEFR did the TBT. That is, a wider range and extent of abilities appear to have been engaged when the items were task-based.

The second research question investigated the way and extent to which the characteristics of GEPT reading test tasks corresponded to those of the TBT tasks derived from the can-do statements in the CEFR and the TLU domain. The characteristics of the texts and items/tasks were also analyzed based on the raters' judgment. Table 5.2 summaries the salient text features for the three tests (GEPT, TBT-social, TBT-work).

Table 5.2: Summary of salient texts features

Texts	Features
GEPT-GRC	B1+, shorter, concrete, expository, educational, straightforward, general (not sociocultural nor context specific), single language function (ideational)
TBT-SRC	B2, longer, some abstract content, argumentative, linguistically and propositionally complex and dense, rhetorical organizations and features more complex and varied, moderately sociocultural specific, single language function (ideational)
TBT-WRC	B2, longer, concrete, linguistically less complex, propositional sparse and complex, fairly sociocultural and context specific, multiple language functions (ideational and manipulative/directive)

As can be seen in Table 5.2, the GRC texts were shorter than the SRC and WRC texts, and they were concrete, expository, educational, straightforward, ideational, and general. The SRC texts were more abstract and argumentative, and the rhetorical organizations and features were more complex and varied. Furthermore, the SRC texts were linguistically and propositionally more complex and dense, not educational, ideational, and sociocultural more specific. The WRC texts were more propositionally sparse and complex, functionally complex, and sociocultural and context specific. Nevertheless, what really set the GRC and the SRC/WRC apart were the differences in "length", "proposition complexity", and "pragmatics". What truly distinguishes them is the amount of complex propositional content that needed to be understood and the multiple language functions and sociocultural specificities associated with these propositions. In other words, the SRC and WRC texts were all much longer and propositionally and pragmatically more complex than the GRC texts.

Summaries of salient item/task features for each subsection across the two tests are provided in Table 5.3. The salient features include operation, explicitness, abstractness, content tested, amount of information, language functions, sociolinguistics, and strategic demands. For the type of operation and amount of information, the percentages of items engaged were reported in parenthesis.

Table 5.3: Summary of salient item/task features

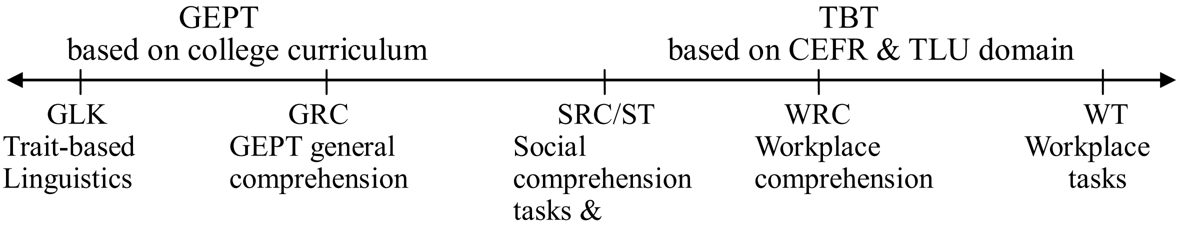
Test	Subsection	Item/task features
GEPT	GLK	To recognize (100%) explicit and concrete lexicon and syntax within or between sentences, no language functions, not sociocultural specific, no strategic demands
	GRC	To recognize(61%) and interpret (39%) explicit, concrete, and straightforward content within a paragraph (47%) and beyond (27%); general, not sociocultural specific, no strategic demands, single language function
TBT	SRC	To recognize (58%) and interpret (42%) explicit, concrete and abstract, and complex content within (17%) and beyond (42%) a paragraph, some sociocultural specificity, limited strategic demands, single language function
	WRC	To recognize (50%) and interpret (50%) explicit, concrete, and complex content within (50%) and beyond (25%) a paragraph, content involved both ideation and manipulative function, required limited strategic demand
	ST	To interpret (83%) complex, somewhat abstract, and implicit content beyond a paragraph (67%) length, input and response are both sociocultural specific, required some strategic demand, simple and frequent linguistic usage
	WT	To interpret (100%) concrete, sparse, and complex content of almost the entire passages (67%), content and responses both involved multiple language functions and are sociocultural and context specific, required extended strategic demand, simple and frequent linguistic usage

As can be seen in Table 5.3, the GLK items/tasks tested test takers' ability to recognize lexicon and syntax in sentences, and the GRC tested test takers' ability to recognize and interpret mostly localized information in the texts. The SRC tested test takers' ability to recognize and interpret paragraph level content that involves sociocultural references, and the WRC tested test takers' ability to recognize and interpret mostly localized information that involves manipulative functions. In addition to testing test takers' ability to interpret explicit and implicit content of more than one paragraph in length that involves sociocultural references and manipulative functions, the ST and WT also tested test takers' ability to strategically convey their understanding of the content in response to situated prompts. Accordingly, the key features that separate the GEPT from the TBT are “operation”, “amount of information” , “language functions”, “sociolinguistics”, and “strategic demands”. What makes them quite different from each other is the amount of input needed to be interpreted rather than recognized and the pragmatic and strategic demands involved when making such interpretations. In other words, the TBT, especially the ST and WT, involves a wider range and extent of characteristics, competences, and input to be processed than the GEPT.

The third research question investigated the way and extent to which the underlying construct of the GEPT reading test corresponds to that of the TBT developed based on the can-do statements in the CEFR and the TLU domain. To address this question, the underlying factor structure of both tests was modeled using confirmatory factor analysis. Based on the analyses of model comparisons and the final model, it became clear that neither the twenty-four variables nor the six subsection factors that comprised the two tests loaded on one single trait. On the contrary, these variables and subsections loaded on their respective distinct subsections and tests, along with a common reading ability that underlay both tests. In other words, these variables and subsections were still distinct in nature even though they all measured a common reading ability. In addition, they have their own varied relationships to the common underlying ability and to each other.

The most distinct subsection factors were the GLK and WT; they had the lowest factor loadings on the common trait, reading comprehension (SMR .72 and .68), and the lowest correlation with each other (SMR .50). These loadings reveal the discrepancies that existed among the three distinct constructs of reading, i.e., linguistic knowledge (GLK), reading comprehension (GRC, SRC, WRC), and situated reading tasks (ST, WT). They also show the divergence between the two distinct approaches, trait-focused and task-focused, to the conceptualizing construct. Furthermore, the GLK was closely related to the GRC (.94) and slightly less related to the SRC (.86). However, the correlations between GLK and the ST (.81), WRC (.78), and WT (.71) were considerably lower. Similarly, the correlations between the GRC and the ST (.85), WRC (.83), and WT (.76) were also considerably lower. These correlations suggest that when the items are more task-based and workplace specific, their association with the GEPT tests decreases, especially the discrete linguistic knowledge. This suggests that there is a continuum of relationships among these reading subconstructs, with the relationships among the trait-based and task-based constructs decreasing as the tasks become more CERF and TLU domain-specific. This continuum is illustrated in Figure 5.1.

Figure 5.1: The relationships between the GEPT and TBT constructs on a continuum



These results are consistent with the results of the CEFR classifications and item/task characteristics analyses, which also support the changing relationships of these subconstructs along a continuum. Table 5.4 summarizes the major differences among these subconstructs based on the characteristics analyses. It also illustrates the varying nature of these constructs and explicates why they differ from each other on the continuum.

Table 5.4: Summary of the major differences among the six constructs

Features	GEPT		TBT-Comprehension		TBT-Situated reading tasks	
	GLK	GRC	SRC	WRC	ST	WT
Language functions	None	Ideational	Ideational	**Input is manipulative,** response is ideational	Ideational	**Input & response are manipulative**
Operation	Only recognize	Mostly recognize	Recognize & interpret	Recognize & interpret	**Mostly interpret**	**Only interpret**
Amount of input	Sentential	Mostly within & a paragraph	One paragraph & beyond	Mostly within & a paragraph	**Beyond paragraph**	**Almost entire passage**
Socio-linguistics	None	Limited	Some	Limited	**Extended**	**Extended**
Strategic demand	None	None	Limited	Limited	Some	**Extended**
Proposition complexity	Straight-forward	Straight-forward	Complex	Complex	Complex	Complex

The qualitative results show that when the items were more task-based and work-specific, not only did they require a significantly higher amount of complex propositions to be interpreted, they also engaged wider a range and extent of language abilities (ideational, functional, and sociolinguistic) and strategic competence. Discrete linguistic knowledge may be relevant to general text reading (SMR with GRC .88 and with SRC .73), but it definitely underrepresents the reading constructs involved in the workplace and situated reading tasks (SMR .60 and .50). The GLK and the WT also appear to differ considerably from the rest of the constructs in the characteristics analyses. The GLK involved no language functions, sociolinguistics, or strategic demand, whereas the WT involved extended manipulate function, sociolinguistics, and strategic demands in both input and response. The GLK only required recognition of sentential input, whereas the WT required interpretation of almost the entire passage. It is noteworthy that the correlation between the GLK and the WRC (.78) is lower than the correlation between the GLK and the ST (.81), as this suggests that the manipulative functions of the WRC seem to have a greater effect on the nature of the reading construct than the ideational types of situated tasks.

For the comprehension constructs (GRC, SRC, and WRC), the characteristics and correlations of the GRC and the SRC seem relatively comparable (SMR .82). Length, proposition complexity, and limited productive responses may affect the difficulty of the tests, but they do not seem to shift the nature of the reading comprehension construct. However, the correlation between the GRC and the WRC is considerably lower (SMR .68), indicating that the manipulative function seems to have affected the nature of the comprehension construct. Furthermore, the GRC has the lowest correlation with the WT (SMR .57), suggesting that the GRC does not accurately represent the reading abilities involved in the workplace. On the other hand, the high correlation between the GRC and the GLK (.94) poses another problem for score interpretation because this correlation is even higher than the correlation between GRC and the SRC. The GRC is intended to measure reading comprehension ability rather than discrete linguistic knowledge; however, the analyses suggest that the GRC is actually more closely related to linguistic knowledge than to the SRC comprehension. This may be due in part to the method effect, as the GLK and GRC both utilize multiple-choice questions. This might also suggest a problem of construct underrepresentation. The GRC does contain a higher proportion of test items measuring recognition of localized information than interpretation of paragraph-length content.

The findings of the CEFR level/category classification, task/text characteristics analyses, and the CFA generally correspond to each other. Table 5.5 summarizes the construct definition of the six subconstructs based on the salient features and their relationships to the CEFR categories and language abilities on the continuum.

Table 5.5: Construct definition of each subsections

Constructs	CEF	Definition	CEFR constructs involved	Language abilities
GLK ↑	--	Can **recognize** syntax and lexicon within and between sentences	---	Grammatical
GRC	B1+	Can **recognize and understand** factual straightforward information in longer texts of common topics	Comprehension	Ideational
SRC	B2	Can **identify, understand, and paraphrase** essentials in long and complex texts dealing with contemporary problems in which the writers adopt particular stances or viewpoints	Comprehension Sociolinguistic Text processing	Ideational Sociolinguistics
ST	B2	Can **synthesize and report** information and arguments in long and complex texts and **give** relevant and appropriate **opinions**	Comprehension, sociolinguistics, text processing, written interaction, info exchange	Ideational Sociolinguistics Strategic
WRC	B2	Can **identify, understand, and paraphrase** essentials in long and complex texts dealing with contemporary problems in which the writers adopt particular stances or viewpoints	Comprehension, Text processing,	Ideational Manipulative
WT ↓	B2	Can **give a description of** how to carry out a **procedure** based on long and complex texts with reasonable precision reliably, using the original text wording and ordering Can **check and confirm** factual routines in long and complex texts and appropriately **advise** on matters related to occupational roles	Comprehension, sociolinguistics, text processing, written interaction, info exchange	Manipulative Sociolinguistics Strategic

As shown in the table, when reading items are more task-based, contextualized, and workplace specific, both the nature and the constituents of the reading comprehension construct shift. The reading abilities involved are broader and more complex. The most complex reading construct, WT, is well represented by the combined effect of comprehension, sociolinguistics, text processing, written interaction, and information exchange. The WT tasks involved the use of linguistics, pragmatics, and strategic competence in order to purposefully and skillfully integrate what was understood from long and complex texts and exchange this understanding with particular interlocutors with particular constrains and conditions in specific contexts. A combination of manipulative function and strategic demand seems to have the greatest effect on the complexity of reading construct.

Both the analyses of ratings and the factor analyses yield similar results. These suggest that while both tests measured reading comprehension, they are quite different in nature. Not only was the TBT more difficult, the constructs engaged are also more complex. The ratings and classifications of the CEFR levels, additional constructs, and text/task characteristics all correspond to the variations identified in the quantitative analyses. In other words, the findings do not support sufficient correspondences existed between the GEPT and TBT in the aspects of CEFR content/levels, task/text characteristics, and underlying constructs. The GEPT construct underrepresents the TLU constructs and its task features. As a result, test takers who pass the GEPT high-intermediate reading test may not be able to read newspapers, magazines, written messages, and work documents at the CEFR level B2 as the LTTC claims it to be. The GEPT test scores do not provide stakeholders with sufficient information regarding the reading ability to be assessed in the TLU domain.

5.2 Implications for language assessment

This study used the GEPT high-intermediate reading test as an example to demonstrate and explicate the issues involved in generalizing test scores to non-test situations. The study identifies and demonstrates quantitatively and qualitatively the way and extent to which two distinct ways of conceptualizing reading constructs, the trait/curriculum-based and the task/domain-based approaches, can lead to divergent construct specifications, difficulty levels, item/text characteristics, and underlying factor structures. When items are more task-based and workplace specific, the less similarity they share with trait/curriculum based test items. The nature and the constituents of the reading comprehension construct shift. Not only do task-based and workplace specific items require a significantly higher amount of complex propositional content to be interpreted rather than recognized, they also demand a wider range and extent of language abilities (ideational, functional, and sociolinguistic) and strategic competence when making such interpretations in relation to context. Among all the combinations of language abilities, that of manipulative function and strategic demand appear to have the most effect on the complexity of reading construct.

The manipulative function and strategic competence constitute the heart of language use in context. The degree of contextualization appears to have a direct and compelling effect on the nature of a reading construct. When an item is contextualized or task-based but the text is not, and the text still involves the same ideational language function as the general reading comprehension construct, the change is small (SMR dropped 10% between the GRC and ST). When the item is not contextualized but the text is and has involved both ideational and manipulative functions, then the change increases slightly (SMR dropped 15% between the GRC and WRC). However, when *both* the item and the text are contextualized, and *both* the input and expected responses involve multiple language functions (ideational, manipulative), the change is amplified (SMR dropped 26% between the GRC and WT). The more the items and texts are contextualized, the more language functions are involved, and the more the language abilities are engaged. The ability to comprehend texts then is different from the ability to comprehend texts in context. The very nature of contextualization changes the nature and constituents of the comprehension construct.

Using Bachman and Palmer's AUA framework, this study strongly suggests that the GEPT is not as meaningful or generalizable as the LTTC claims it is. GEPT test scores do not provide stakeholders with sufficient information about the ability to be assessed in the TLU domain, and the GEPT tasks do not have a sufficient degree of correspondence to the TLU tasks. For instance, the GLK involves no language functions, sociolinguistics, or strategic demand, while the WT involves extended manipulate functions, sociolinguistics, and strategic demands in both input and response. The GLK only requires recognition of sentential input, whereas the WT requires interpretation of nearly the entire passage. The WT tasks involve the use of linguistics, pragmatics, and strategic competence in order to purposefully and skillfully integrate what has been understood from long, complex texts and to exchange this understanding with particular interlocutors under particular constrains and conditions in specific contexts. Little of these skills and capacities are captured in the GEPT's GLK and GRC multiple-choice questions. In Bachman's (1990) terms, the GEPT high-intermediate reading test does not define language ability in such a way that the test methods they use elicit language test performance that is characteristic of language performance in non-test situations. In other words, due to inadequate sampling of the target constructs and its task characteristics, GEPT test scores do not appear to generalize to performance in the target domain.

5.3 Limitations

The study was limited in that the sample size collected for the study was only sufficient for one group analysis, and was too small for conducting a cross-validation study using separate analyses of subgroups. This is because a CFA of 24 parcel variables in the test would have required a much larger data set in order to generate accurate cross-validation results. This study was also limited in the number of raters recruited and the number of the GEPT test forms that were analyzed and administered. Further studies might test the models using different and larger sets of test takers, raters, and test forms to collect more evidence to investigate the meaningfulness and generalizability of the test. Furthermore, other types of analyses, such as model testing through item response theory and verbal protocol analysis could be used to investigate other possible sources of variation to account for the large and unexplained variances in these factor models.

Appendix A: Salient characteristics of Reading in the CEFR Manual

Revised Table A2. Salient Characteristics: Reading

	Topic	Text features	Restrictions	Process	What is tested
C1	• Abstract and complex topics • Social, academic and professional life	• Lengthy, complex texts of various kinds • A wide range of idiomatic expressions and colloquialisms, metaphors, figurative language • A wide range of infrequent, specialized, or technical words • Register shifts (mixture of formal, causal) • Implied attitudes and relationships • Factual and literary • Technical	May occasionally need to: • confirm details (with dictionary) if outside field • re-read difficult sections	• Understand • Identify	• Finer points of detail • Implied as well as stated opinions • A wide range of idiomatic expressions and colloquialisms • Register shifts • Implied attitudes and relationships
B2+	• A wide range of familiar and unfamiliar topics • Social, academic and professional life		• Standard, non-idiomatic: • May occasionally need to confirm details (with dictionary) if outside field	• Understand	
B2	• Reasonably familiar concrete and abstract topics • Related to field of interest/specialty	• Long and complex texts • **News items, articles and reports** • Deal with contemporary problems in which the writers adopt particular stances or viewpoints • Propositionally and linguistically complex text • Literary prose	• **Standard** • Clearly signposted/ signaled with explicit markers • If can re-read difficult sections	• Scan quickly • Identify	• Relevant details • Content and relevance • Specific details
				• **Understand** (with a large degree of independence)	• **Main ideas,** opinions • Essentials/essential meaning • **Complex lines of argument** • **Speaker/writer mood, tone**
B1+	• Common everyday or job-related topics	• Longer texts • Different texts, different parts of a text • Argumentative text	• Standard • **Straightforward** • Clearly signposted/ signaled with explicit markers	• Scan • Locate	• Desired information
				• Understand	• Straightforward factual information content • General message • Main conclusions • Specific details
B1	• Familiar topics • Regularly encountered in a school, work or leisure context	• **Straightforward** newspaper articles • Straightforward **factual** texts • Clearly signaled argumentative texts • Everyday materials**: letters, brochures, short official documents** • Short narratives • Descriptions of events, feelings, wishes • Detailed directions • Simple technical information e.g. operating instructions	• Clear • Standard • **Straightforward** • High frequency everyday language	• Scan • Find and understand	• Desired information • Relevant information
				• **Identify** • **Recognize**	• Main conclusions • Line of argument (not in detail) • Significant points

Appendix B: Task characteristic rating instrument

Analyzing Items/tasks	
Characteristics	Description
1.Input linguistics Lexicon Syntax	1=only frequent vocab 1=only simple structure 2=mostly frequent vocab 2=mostly simple structure 3=some infrequent vocab 3=some range of complex structure 4=wide range infrequent vocab 4=wide range of complex structure
2. Content tested	Identify the content tested (e.g. main idea, details, inference about main idea/details, paraphrase/summarize of main idea/details, opinion forming, or others)
3.Implicitness	1= only explicit 2= mostly explicit 3= somewhat implicit 4= mostly implicit
4. Abstractness	1= only concrete content 2= mostly concrete content 3= relatively abstract content 4= mainly abstract content
5. Operation	1= Recall: recognize/retrieve a specific piece of information alone 2= Concept: infer/interpret/compare/summarize meaning 3= Reasoning: analyze/evaluate/justify/generalize/synthesize
6. Amount of input must be processed to answer an item correctly	1= requires only localized understanding of 1-2 sentences 2 = requires an understanding of one paragraph 3= requires an understanding of more than one paragraph 4= requires an understanding of the entire passage
7. Language functions in the input and response	Identify language functions involved (i.e. ideational=I, manipulative=M, heuristic=H, imaginative=IM)
8. Sociolinguistics	Label the items based on the CEFR sociolinguistics levels
9. Strategic demand when producing a response	The extent and amount of strategic assessment and planning required in order to produce the answer: 0= none, 1=limited, 2=some, 3=extended, 4= demanding
10. Response linguistics Lexicon Syntax	1=only frequent vocab 1=only simple structure 2=mostly frequent vocab 2=mostly simple structure 3=some infrequent vocab 3=some range of complex structure 4=wide range infrequent vocab 4=wide range of complex structure

Appendix B-continue

Analyzing Texts	
Characteristics	Description
1.Domain	Personal(social), public, occupational, OR educational
2. Discourse type	Descriptive(D), narrative(N), expository(E), argumentative(A), AND/OR instructive(I) Use the following five descriptors to describe these variables, to rate the degree to which the feature, type of reasoning, or organization appears to be: 0 = no use 1 = present to a very limited extent, but not used to organize the overall structure of the passage in any meaningful way 2 = present, but only used to organize the structure of the passage in minor ways/to a minor extent, or to organize a particular portion of the passage 3 = used to a major extent in organizing the passage 4 = used as the primary or sole means of organizing the structure of the passage
3.Abstractness	1= content is highly concrete 2= content is relatively concrete 3= content is relatively abstract 4= content is highly abstract
4. Linguistics Lexicon Syntax	1=only frequent vocab 1=only simple structure 2=mostly frequent vocab 2=mostly simple structure 3=some infrequent vocab 3=some range of complex structure 4=wide range infrequent vocab 4=wide range of complex structure
5. Rhetorical organization Features	Perceived complexity: Identify major types of features used: 1= very simple examples, contrast, cause/effect, 2= moderately simple problem/solution, classification, analysis 3= moderately complex 1=limited, 2=present, minor/segment 4= very complex 3=major, 4=primary
6. Proposition density and complexity	1= highly sparse 1= highly straightforward 2= moderately sparse 2= moderately straightforward 3= moderately dense 3= moderately complex 4= highly dense 4= highly complex
7. Pragmatics Directness Speech act Language function	1= indirect, 2=relatively indirect, 3=relatively direct, 4=highly direct 1=very small number, 2=relatively small, 3=relatively large, 4=large Identify language functions involved (ideational, manipulative, heuristic, imaginative)
8. Sociolinguistics Specificity Language use Formality	1=little, 2=somewhat, 3=moderately, 4=highly specific culture/context Identify any use of idioms, slangs, dialects, and figurative language Register: 1=intimate, 2=casual, 3=in between, 4=formal
9. Amount of input	Number of words and sentences in texts and questions.

REFERENCES

Alderson (2000). *Assessing Reading*. Cambridge University Press.

Alderson, J. C., Figueras, N., Nold, G., Takala, S., & Tardieu, C. (2006). Analysing Tests of Reading and Listening in Relation to the Common European Framework of Reference: The Experience of The Dutch CEFR Construct Project. *Language Assessment Quarterly*, 3(1), 3-30.

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

Bachman, L. F. & Palmer, A. S. (1996). *Language testing in practice: designing and developing useful language tests*. Oxford: Oxford University Press.

Bachman, L. F. (2002a). Some reflections on task-based language performance assessment. *Language Testing*, 19(4), 453-476.

Bachman, L. F. (2002b). Alternative interpretations of alternative assessments: Some validity issues in educational performance assessments. *Educational measurement: Issues and Practice*, 21(3), 5-18.

Bachman, L. F. (2005). Building and supporting a case for test use. *Language Assessment Quarterly, 2,* 1-34.

Bachman, L. F. (2007). What is the construct? The dialectic of abilities and contexts in defining constructs in language assessment. *Language Testing Reconsidered*. Fox, J., Wesche, M., Bayliss, D., Cheng, L., Turner, C. E., & Doe, C., editors. Ottawa: University of Ottawa Press.

Bachman, L. F. & Palmer, A. S. (1981). The construct validation of the FSI oral interview. *Language Learning*, 31, 67-86.

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

Bentler, P. M. (1990). Comparative fit indexes in structural models. *Psychological Bulletin, 107(2),* 238-246.

Bentler, P. M. (2006). *EQS 6 structural equations program manual*. Encino, CA: Multivariate Software, Inc.

Bentler, P. M., & Wu, E.J. C. (2003). *EQS 6.1 for Windows.*

Brown, J. D., Hudson, T. D., Norris, J. M. & Bonk, W. (2002). *An investigation of second language task-based performance assessments.* Honolulu, HI: University of Hawaii Press.

Brown, J. D. (2004). Performance assessment: existing literature and direction for research. *Second Language Studies, 22*(2), 91-139.

Browne, M. W. & Cudeck, R. (1993). Alternative ways of assessing model fit. In: Bollen, K. A. & Long, J. S. (Eds.) *Testing Structural Equation Models*. pp. 136–162. Beverly Hills, CA: Sage.

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

Carroll, J. B. (1968). The psychology of language testing. In Davies, A., editor, *Language testing symposium: a psycholinguistic approach*. London: Oxford University Press, 46-69.

Canale, M. & Swain, M. (1980). Theoretical bases of communicative approaches to second language testing and testing. *Applied Linguistics* 1, 1-47.

Canale, M. (1983). On some dimensions of language proficiency. Oller, J, editor, *Issues in Language Testing Research*, Newbury House, Rowley, MA, 333-342.

Chapelle, C. (1998). Construct definition and validity inquiry in SLA research. In Bachman, L. F.& Cohen, A. D., editors, *Interfaces between second language acquisition and language testing research.* New York: Cambridge University Press, 32-70.

Chalhoub-Deville, M. (2003). Second language interaction: Current perspectives and future trends. *Language Testing* 20(4), 269-383.

Clark, J. L. D. (1975). Theoretical and technical considerations in oral proficiency testing. In Jones, R. L. & Spolsky, B., editors, *Testing language proficiency.* Arlington, VA: Center for Applied Linguistics, 10-24.

Council of Europe. (2001). *Common European Framework of Reference for Languages: learning, teaching, assessment.* Cambridge: Cambridge University Press.

Council of Europe. (2003). *Relating language examinations to the Common European Framework of Reference for Languages: learning, teaching, assessment. A Manual.* Strasbourg, France: Council of Europe.

Council of Europe. (2005). *Reading and listening items.* Strasbourg, France: Council of Europe. Compact disc.

Douglas, D. (2000). *Assessing language for specific purposes: theory and practice*. Cambridge: Cambridge University Press.

Hu, L.T., & Bentler, P.M. (1998). Fit indices in covariance structure modeling: sensitivity to underparameterized model misspecification. *Psychological Methods, 3,* 424 –

Jones, R. L. (1985). Second language performance testing: An overview. In P. C. Hauptman, R. LeBlanc, & M. B. Wesche (Eds.), *Second language performance testing* (pp. 15-24). Ottawa: University of Ottawa Press.

Kane, M. (1992). An argument-based approach to validity. *Psychological Bulletin*, 112-3, 527-535.

Kane, M. (2001). Current concerns in validity theory. *Journal of Educational Measurement, 38,* 319-342.

Kane, M., Crooks, T., & Cohen, A. (1999). Validating measures of performance. *Educational measurement: Issues and practice, 18,* 5-17.

Kline, R. B. (1998). Principles and practice of structural equation modeling. New York: *Guilford Press.*

Kunnan, A., & Wu, J. (2009). The Language Training and Testing Center, Taiwan. Cheng, L. and Curtis, A., editors, *English Language Assessment and the Chinese Learner*, Routledge.

Language Training and Testing Center [LTTC]. (2000, September). *GEPT High-Intermediate Level Research Report*. Taipei, Taiwan: Author. http://www.lttc.ntu.edu.tw/research/hi/中高級研究報告.pdf

Long, M. H. & Norris, J. M. (2000). Task-based language teaching and assessment. In Byram, M., editor, *Encyclopedia of language teaching.* London: Routledge, 597-603.

Ma,T. M. & Li, S. F (2009). *Bridging Test Construct and Beneficial Washback Effects: Revising the GEPT High-intermediate Reading Test*. LTTC. Taipei, Taiwan. http://www.lttc.ntu.edu.tw/academics/Bridging_Test_Construct_and_Beneficial_Washback_Effects.pdf

Mislevy, R. J., Steinberg, L. S., & Almond, R. G. (2003). On the structure of educational assessments. Measurement: Interdisciplinary Research and Perspectives, 1(1), 3-62.

Norris, J. M., Brown, J. D., Hudson, T., & Yoshioka, J. (1998). *Designing second language performance assessments.* (Vol. SLTCC Technical Report #18). Honolulu: Second Language Teaching and Curriculum Center, University of Hawaii at Manoa.

Purpura, J. (2007). Assessing communicative language ability: Models and their components. *Encyclopedia of language teaching.* London: Routledge,

Skehan, P. (1998). *A cognitive approach to language learning*. Oxford: Oxford University Press.

SPSS Inc. (2008). *SPSS for Windows release 16.0 standard version*. Chicago, IL: SPSS Inc.

Toulmin, S. E. (2003). *The uses of argument* (Updated ed.). New York: Cambridge University Press.

Ullman, J. B. (2001). Structural equation modeling. In B. G. Tabachnick & L. S. Fidell (Eds). *Using Multivariate Statistics* (4th ed.). Boston: Allyn and Bacon.

Upshur, J. A. (1979). Functional proficiency theory and a research role for language tests. In Briere, E., & Hinofotis, F. B., editors, *Concepts in language testing: some recent studies.* Washington, DC: TESOL, 75-100.

Wu, J. (2002). Assessing English proficiency at advanced level: The case of the GEPT. *Proceedings of the International Conference on Language Testing and Language Teaching*, 93-100. Shanghai, China.

Wu, J. R. W. & Wu, R. Y. F. (2010). Relating the GEPT reading comprehension tests to the CEFR. *Aligning Tests with the CEFR: Reflections on using the Council of Europe's draft Manual.* Martyniuk, W. (ed). Cambridge: University of Cambridge Press.