A COMPARISON OF SUBSCORE REPORTING METHODS FOR A STATE ASSESSMENT OF ENGLISH LANGUAGE PROFICIENCY

A COMPARISON OF SUBSCORE REPORTING METHODS FOR A STATE ASSESSMENT OF ENGLISH LANGUAGE PROFICIENCY

By

Tanya Longabach

Submitted to the graduate degree program in the Department of Psychology and Research in Education and the Graduate Faculty of the University of State in partial fulfillment of the requirements for the degree of Doctor of Philosophy.

_____

Chairperson: Dr. Vicki Peyton

_____

Co-chairperson: Dr. William Skorupski

_____

Dr. Neal Kingston

_____

Dr. Bruce Frey

_____

Dr. Lizette Peter

Date Defended: April 1, 2015

The Dissertation Committee for Tanya Longabach certifies that this is the approved version of the following dissertation:

A COMPARISON OF SUBSCORE REPORTING METHODS FOR A STATE ASSESSMENT OF ENGLISH LANGUAGE PROFICIENCY

_____

Chairperson: Dr. Vicki Peyton

Co-Chairperson: Dr. William Skorupski

Date approved: April 1, 2015

ABSTRACT

Educational tests that assess multiple content domains related to varying degrees often have subsections based on these content domains; scores assigned to these subsections are commonly known as subscores. Testing programs face increasing customer demands for the reporting of subscores in addition to the total test scores in today's accountability-oriented educational environment. While reporting subscores can provide much-needed information for teachers, administrators, and students about proficiency in the test domains, one of the major drawbacks of subscore reporting includes their lower reliability as compared to the test as a whole. This dissertation explored several methods of assigning subscores to the four domains of an English language proficiency test (listening, reading, writing, and speaking), including classical test theory (CTT)-based number correct, unidimensional item response theory (UIRT), augmented item response theory (A-IRT), and multidimensional item response theory (MIRT), and compared the reliability and precision of these different methods across language domains and grade bands. CTT and UIRT methods were found to have similar reliability and precision that was lower than that of augmented IRT and MIRT methods. The reliability of augmented IRT and MIRT was found to be comparable for most domains and grade bands. The policy implications and limitations of this study, as well as directions for further research, were discussed.

Acknowledgements

I would like to extend my sincere gratitude to my advisor, Dr. Vicki Peyton, for her patience and faith in my abilities. I appreciate the help and encouragement, and most insightful comments, of all the members of my committee: Dr. William Skorupski, Dr. Bruce Frey, Dr. Neal Kingston, and Dr. Lizette Peter. I appreciate Dr. Peter's invaluable input from the perspective of second language instruction.

My many thanks go out to Dr. Jonathan Templin, who answered my numerous questions. I am also grateful to Dr. John Poggio's participation in my academic career. I am grateful for my family's patience with me and my academic endeavors. Last but not least, my thanks go out to my friends, many of whom never went to college, and my classmates, who walked with me every step of this way. You knew I could do it even before I did. But I couldn't have done it without you.

Table of Contents

List of Tables

List of Figures

A Comparison of Subscore Reporting methods for a State Assessment of English Language Proficiency

CHAPTER 1

INTRODUCTION

Testing programs face increasing customer demands for the reporting of subscores in addition to the total test scores in today's accountability-oriented educational environment. Educational tests often have subsections based on content categories; scores assigned to these subsections are commonly known as subscores (Haberman, Sinharay, & Puhan, 2009). According to Haladyna and Kramer (2004), an interest in getting information about subscores is increasing in the environment of high stakes testing due to the fact that students want to know their strengths and weaknesses to remediate the latter, and teachers would like to be able to better evaluate their performance and focus on the areas of possible instructional improvement. Testing programs in general face higher demands for subscores to provide more detailed information about examinees than that provided by the total score (Haberman, 2008). Brennan (2012) notes that users of test scores often demand that subscores be reported in addition to the total test scores for diagnostic purposes.

Different methods of subscore reporting have been investigated, including those based on classical test theory (CTT) vs. item response theory (IRT) framework, including unidimensional item response theory (UIRT) vs. multidimensional item response theory (MIRT) within the IRT framework, and augmented vs. non-augmented methods. Since subscore information can play an important role in student evaluation and curriculum planning, it is essential that methods of subscore reporting that deliver the most reliable and informative results are used.

Importance of Subscores in Educational Testing

The No Child Left Behind Act of 2001 (NCLB, 2001) requires statewide testing programs to report diagnostic information for examinees that allows parents, teachers, and administrators to understand and address the specific academic needs of students, and to include information regarding achievement on academic assessments aligned with State academic achievement standards (Skorupski, 2008). Luecht (2003) notes that there are a number of positive reasons to report subscores. For example, failing candidates understandably may want to know how they did on particular parts of the test that may have contributed to their failure to help them study for a retest. Teachers often use individual or aggregate test results from their students to gauge the success of their curricula and make necessary modifications. Ling (2009) states that reporting subscores may provide test-takers and test users with better knowledge of test performance on a subset of items, especially when the sub-domains of a test vary by content or underlying construct. In addition to the total test score, students or teachers may also be interested in knowing examinees' competence in each aspect of the curriculum.

Assessment plays a central role in the education of English language learners (ELLs) and bilingual children. NCLB (2001) uses the term Limited English Proficient (LEP) to describe ELL students; from this point forward these terms will therefore be used interchangeably. NCLB defines an ELL student as an individual who

a) is 3 to 21 years old;

b) is enrolled or preparing to enroll in elementary or secondary school;

c) was not born in the U.S. or whose native language is not English;

d) is a Native American, Alaskan Native, or a resident of outlying areas;

e) comes from an environment in which a language other than English has had a significant impact on an individual's English language proficiency;
f) is migratory and comes from an environment where English is not the dominant language; and
g) has difficulties in speaking, reading, writing, or understanding the English language that may deny the individual the ability to meet the state's proficient level of achievement, to successfully achieve in classrooms where English is the language of instruction, or to participate fully in society.

Teachers generally use assessments to monitor language development in students' first or second language and track the quality of their day-to-day subject matter learning (Hakuta & August, 1997). In addition, these tests are frequently used to assign ELLs to specific instructional services in schools and for reclassifying students as proficient once they developed sufficient language skills and are deemed ready to exit English as a second language (ESL) services. Teachers and administrators use information from total proficiency test scores as well as specific domain scores to identify ELL students, determine their eligibility for placement in specific language programs, and monitor their progress in and readiness to exit from these programs (Crawford, 2004; Hakuta & Beatty, 2000). A student misclassified as ELL or as ELL of an inappropriate ability level may be assigned a curriculum that does not correspond to his or her proficiency level and thus receive inappropriate instruction. Alternately, inappropriate instruction may result in low performance, which may in turn result in misclassification. Invalid assessment may result in misclassification, and consequently in inappropriate instruction. Valid assessments, however, may provide diagnostic information that can inform instruction and classification and are therefore an integral part of ELL instruction (Abedi, 2004). It is clear that all three components – assessment, instruction, and classification – inform and impact each other in providing optimal education for ELL students (figure 1).

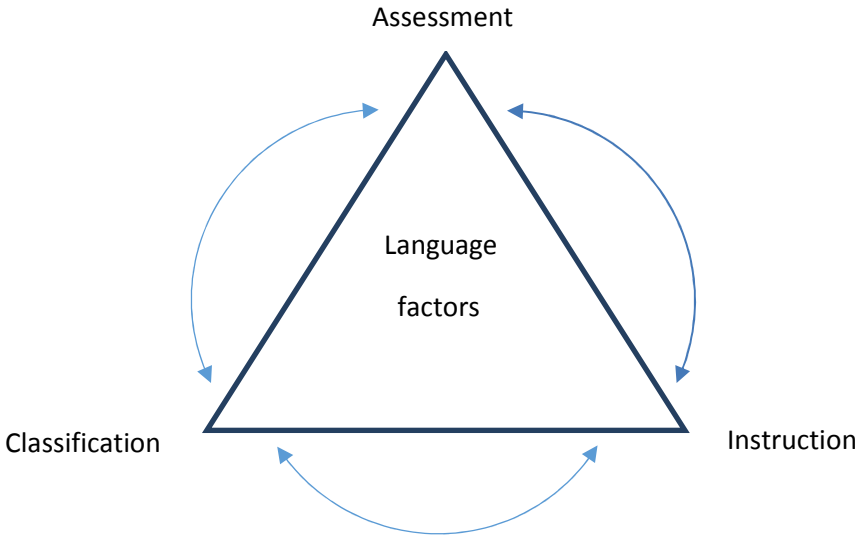

Figure 1. Interactive school achievement model for ELL students (Abedi, 2004, p.12).

Need for Distinct and Reliable Subscores

Providing subscores that have value is, therefore, essential for students in general, and for ELL population in particular. While subscores that are valid and reliable can be invaluable in the educational environment, lack of these qualities may lead to inaccurate decisions and other undesirable consequences. According to the American Educational Research Association (AERA) standards (2014), it is essential to have solid justification for reporting of subscores, including confirming their reliability and validity:

> Standard 1.14: When interpretation of subscores, score differences, or profiles is suggested, the rationale and relevant evidence in support of such interpretation should be provided. Where composite scores are developed, the basis and rationale for arriving at the composites should be given. (p. 27)

> Standard 6.12: When group-level information is obtained by aggregating the results of partial tests taken by individuals, evidence of validity and reliability/ precision should be reported for the level of aggregation at which results are reported. Scores should not be reported for individuals without appropriate evidence to support the interpretations for intended uses (p. 119).

When subscores influence instructional decisions, the relationship between the two should be made transparent:

> Standard 12.19: In educational settings, when score reports include recommendations for instructional intervention or are linked to recommended plans or materials for instruction, a rationale for and evidence to support these recommendations should be provided (p.201).
>
> Comment: … when the patterns of subscores on a test is used to assign students to particular instructional interventions, it is important to provide both a rationale and empirical evidence to support the claim that these assignments are appropriate (p.201).

Two main issues with subscore reporting have been identified: their lack of reliability and lack of distinctiveness from the total score (Haberman, 2008; Haberman & Sinharay, 2008; Haberman et al., 2009; Sinharay et al., 2011). Low reliability of the subscores is usually caused by the fact that subscales for which these subscores are reported consist of a very small number of items. Therefore the reliability of such a subscale is far less than the reliability of the total test, and the subscore is influenced by random error more than by the actual student performance (Haberman, 2008; Sinharay et al., 2011).

The issue of distinctiveness has to do with the amount of distinct information provided by the smaller subset of items used for subscores that is not reflected in the total test score

(Haberman, 2008). When subscores essentially replicate the information provided by the total test score and do not add any information about the student abilities above and beyond the total score, they are considered uninformative. Haberman (2008) notes that the problem of subscore distinctiveness reflects the nature of the test, rather than the method used to develop a subscore; tests that are not built to yield diagnostic (subscore) information tend not to do so. If the structure of a construct is unidimensional, in other words, it only has one subscale, attempts to report subscores as if it is a multidimensional construct will usually result in subscores that are not distinct from the total score.

Statement of the Problem

Title III of NCLB requires assessment of ELL students' English proficiency on an annual basis and provides supports to states to develop their own reliable and valid measures of students' proficiency. Specifically, Section 3102 (8) of the Title III of NCLB 2001 (titled *English Language Acquisition, Language Enhancement, and Academic Achievement Act*), states that the purpose of this legislature is to hold state educational agencies, local educational agencies, and schools accountable for increases in English proficiency and core academic content knowledge of ELL students by requiring demonstrated improvements in the English proficiency in the course of each fiscal year, and adequate yearly progress (AYP).

Attention has been drawn to ELL achievement gaps and to the need for instruction tailored to the ELL specific needs, appropriate and fair testing of ELLs, early identification of problems they may have and the obstacles to their academic success, and better ability to account for ELL achievement and progress. There is an increasing demand for language proficiency assessments that will provide meaningful information educators can use to tailor instruction to ELLs' linguistic and academic needs, thus improving outcomes and narrowing achievement gaps

for these students. Since English language proficiency does not develop uniformly in all domains – listening, reading, speaking, and writing (Jamieson et al., 2000; Chapelle et al., 2011), it is helpful for teachers to know which specific domain abilities may be lagging behind and which abilities constitute strengths for specific students to provide optimal instruction tailored to individual student needs.

A state assessment of English language proficiency reviewed in this study was therefore developed to address the requirements of NCLB. The assessment was designed to:
(1) measure specific indicators within the state Curricular Standards for English to Speakers of Other Languages (ESOL) to indicate a student's level of proficiency with the English language in reading, writing, listening and speaking;
(2) produce data that capture trends across the state and can measure progress in meeting Annual Measurable Achievement Objectives (AMAOs) for Title III accountability requirements; and
(3) provide data on which to base decisions about designing instruction for English Language Learners (ELLs) (Peyton et al., 2009).

The state provides subscore reports and the total score reports to all the stakeholders, including parents of ELL students. Therefore it is essential that subscores are reported in a manner that ensures reliability, precision, and fairness for all ELL students. Currently, the assessment uses unidimensional IRT subscore calculation methods (Peyton et al., 2009). This method, while it represents an improvement in subscore and total score reporting methods in comparison to those based on the classical test theory, still may not be the most reliable and precise method of subscore reporting (Skorupski, 2008; Edwards & Vevea, 2006; Yao, 2010). One reason UIRT subscore reliability may not be optimal is because it does not take into consideration correlations between subscales, which can serve as a source of additional information about student domain abilities. MIRT and subscore augmentation methods, on the

other hand, are able to use this additional subscore correlation information to make subscores more reliable. Due to the nature of the construct being tested in a language proficiency assessment, the subscores in different language domains are usually correlated, indicating that there is a relationship between the domains of listening, speaking, reading, and writing, and the subscore estimation methods that use the correlation between domains can take advantage of this relationship. Also, since language proficiency does not develop uniformly, these correlations change with the students' age and proficiency levels (Alderson, 2007; Chapelle et al., 2011). For the specific methods of subscore reporting that take advantage of these correlations between domains, these correlation changes may impact subscore reliability. The same processes that affect subscore reliability also impact the standard error of measurement at specific score levels (known as conditional SEM, or CSEM). The fact that domain score values may have different CSEM at the extremes and in the middle of ability distribution have been brought up both in the CTT and IRT frameworks (Green et al., 1984; Hambleton & Swaminathan, 1985). Varying CSEM values in different domains at different ability levels directly affect fairness of assessment for students with different degrees of proficiency, as well as subscore reporting precision. The more reliable subscore estimation methods also tend to have lower CSEM, and therefore be more precise in the assignment of a score for an examinee.

Purpose of the Study

A number of studies have been conducted to examine different methods of subscore reporting, such as those comparing augmented vs. non-augmented methods of subscore reporting (Skorupski, 2008; Wainer et al., 2001; Puhan et al., 2010; Sinharay, 2010; Skorupski and Carvajal, 2010; Stone et al., 2010; de la Torre, Song, & Hong, 2011), those comparing CTT vs.

IRT methods (Luecht, 2003; Bock, Thissen, & Zimowski, 1997; Yao & Boughton, 2007; Shin, 2007; Dwyer et al., 2006; Haberman & Sinharay, 2010; Thissen & Orlando, 2001), and those comparing UIRT and MIRT methods (Luecht, 2003; Boughton, Yao, & Lewis, 2006; Yao & Boughton, 2007; Cheng, Wang, & Ho, 2009; Wang, Chen, & Cheng, 2004). However, there are several differences between the present study and the ones conducted previously. First, many previous studies use simulated data as the main source of determining which method of subscore reporting has highest reliability and lowest SEM. The main advantage of using simulated data is that the item and person parameters are known. Additionally, the correlations between subscores are usually known, preset, and constant across all examinees, grades, and ability levels (Boughton, Yao, & Lewis, 2006; Yao & Boughton, 2007; De la Torre, Song, & Hong 2011; De la Torre & Patz, 2005; Edwards & Vevea, 2006). This uniformity of correlations, however, is unlikely to occur in real data. Also, the data are generated based on the assumption of a model that the data should follow, which also may not always be the case with real data (De la Torre & Song, 2009). In real data, correlations between subscores usually vary across grades and ability levels. While the true examinee parameters are not known in the real data, it allows for the examination of irregularities brought on by real life situations. Previous studies suggest that all simulated studies be checked with real data (Luecht, 2003; Boughton, Yao, & Lewis, 2006; de la Torre & Song, 2009; DeMars, 2005; Dwyer et al., 2006). DeMars (2005) suggests that while the conclusions based on a real dataset are likely to generalize to at least other datasets with similar characteristics, the conclusions based a simulated dataset may not generalize to any real dataset at all.

Second, while a number of studies examining different methods of subscore calculation are available, to our knowledge there have been no previous studies conducted specifically on English language proficiency assessment subscore calculation. Due to the fact that constructs

being tested may be structured differently (e.g., mathematics, science, language arts), and specifically correlations between domains may vary across grades and ability levels, it is advisable that many different types of tests' methods of subscore reporting are examined due to these differences (De la Torre, Song, & Hong, 2012).

Third, while the purpose of previous studies of subscore reporting methods was to determine which methods have the highest overall reliability, these studies usually do not attempt to determine the differences between CSEM in different domains at different ability levels across grades. Since determining and reporting CSEM for different ability levels is encouraged by the Standards for reporting, and in general impacts fairness of a test, it is important that CSEM values for subscore reporting methods are compared.

Finally, very few studies that examined the reliability of different methods of subscore reporting used tests that combined dichotomous (scored as pass or fail) and polytomous (scored in several categories) items (Wang, Chen, & Cheng, 2004); consequently, tests scored with the generalized partial credit model have not been used for subscore method comparison. Our study addresses these gaps in knowledge as related to English language proficiency assessment.

Research Questions

In this study, the following research questions will be addressed:

1. How do the four methods of subscore reporting for four language domains compare in terms of subscore reliability and subscore correlations across five grade bands (0-1, 2-3, 4-5, 6-8, and 9-12)?

2. How do the four methods of subscore reporting for four language domains compare in terms of subscore precision across five grade bands?
3. How do the reliability and precision of these four methods of subscore reporting in the four language domains compare in the estimation of ability of examinees within five grade bands at different proficiency levels?

The main purpose of this study was to compare the reliability of several methods of subscore reporting an English language proficiency assessment. As is commonly the case, the state English language proficiency assessment consists of four domains: listening, speaking, reading, and writing. We compared the following methods: raw CTT subscore reporting, non-augmented IRT reporting (which is the present method of subscore reporting); augmented IRT method (IRT-R); and MIRT reporting. The first two methods are non-augmented, that is, they do not use information from other sources, such as individual or group scores, or correlations between subscores. The second two methods use data augmentation, that is, they use data from other sources to augment the subscores. We compared the subscore reporting methods based on their reliability and precision (standard error of measurement (SEM), and conditional standard error of measurement (CSEM) related to examinee proficiency levels.

Summary and Significance of the Study

In summary, this study compared subscore reliability, precision, and correlations between four methods of subscore reporting on an English language proficiency assessment (raw CTT subscore reporting; IRT reporting (which is the present method of subscore reporting); augmented IRT (IRT-R) method; and MIRT reporting overall and across different grade bands and ability levels. The study's significance is in its potential ability to impact the accuracy and fairness of subscore reporting on an English language proficiency assessment. This issue is

specifically important to ELL students, whose placement and instruction depends on the accurate determination of proficiency in each of the four domains of language proficiency assessment. In addition, this study provides some information on how the variation in the correlational structure between the domains of a construct across grade bands and ability levels may impact subscore reporting.

In Chapter 2, we reviewed literature pertinent to the concepts of reliability and measurement precision under the CTT and IRT frameworks, including different methods of measuring reliability, standard error of measurement, and conditional standard error of measurement. We also reviewed different methods of subscore reporting under the CTT and IRT frameworks, as well as the augmented and non-augmented methods of subscore reporting. We described the present studies comparing several methods of subscore reporting. In addition, we discussed some issues pertaining to the correlation of domains in language proficiency testing.

In Chapter 3, we discussed the methods used to compare subscore reporting techniques from the perspective of reliability, SEM and CSEM. Chapter 4 describes the results from the data analyses, and Chapter 5 presents discussion and future directions based on the findings of this study.

CHAPTER 2

LITERATURE REVIEW

The two main issues with subscore reporting are their reliability and distinctiveness. The lower reliability of subscores as compared to the total score stems from a smaller number of items on a subscale, while the lack of distinctiveness usually has to do with the unidimensional structure of the construct that is being tested. We will limit our discussion to the topic of subscore reliability, since that is the focus of this study. In this chapter we will first outline some issues specific to language proficiency testing, such as defining the concept of language proficiency and deciding what language domains to include in the testing. We will then discuss current approaches to methods of subscore reporting that increase reliability within CTT and IRT frameworks, and review studies that compared several methods of subscore reporting with respect to subscore reliability. We will also examine how CTT and IRT treat the concept of reliability, and how reliability, SEM, and CSEM are measured in these frameworks.

General Issues Related to Language Proficiency Testing

Many issues related to validity and reliability surround language proficiency testing. Among the most significant ones are lack of clarity regarding the definitions and boundaries of the constructs being tested; lack of certainty about how different linguistic domains are related and how growth or weaknesses in one domain impacts other domains; the appropriate tasks to assess language domains; difference in opinions regarding how assessment should be conducted between professionals in linguistics and educational measurement; unavailability of clear-cut definitions of language proficiency; and issues of social justice surrounding language proficiency testing.

Construct Definition

Language learning is unique among school subjects in the range of learner attributes which it engages, including cognitive, psychomotor, and affective attributes (Coleman, 2004). There is far more to it than intellectual understanding (Jones, 2011). One goal of various professionals involved in language proficiency testing is to make explicit the claims about the meaning of test scores, and the rationales and evidence that can support these claims. Careful construct definition followed by the construction and validation of methodologies to collect data about language proficiency is essential to the support of test validity. According to Chapelle (2011), critical language testing questions would include the following: "What are the assumptions about language underlying the test? Should language be measured in the manner that it is in the test? What are the effects of measuring such a construct in this way? Who benefits and who loses when this test is used as proposed?" (p.24).

Kane (2011) also draws attention to the elusive qualities of the construct of language proficiency by saying that we may think what we know what language proficiency means, but identifying the range of performances to include in the target domain may be complicated once we attempt to specify the domain well enough to use it as a guide to developing an assessment plan. These issues can get progressively more complex as we move from the mechanics of language use to the general concept of communicative competence, or to the concept of communicative competence for specific purposes (e.g., academic language vs. interpersonal communication). As the domain is being specified, it is important to make the boundaries of the domain clear. Spolsky (2007) points out that it is hard to come up with a short and explicit answer to the question, "What does it mean to know a language?" While one may say that knowing a language means to use it like a native speaker, the next issue may be the uncertainty

of the concept of a "native speaker" and the questionable idea of holding the native speaker as the gold standard to be aimed at by a language learner.

Alderson (2007) argues that in order to diagnose language development, we need to have a clear idea of how foreign language ability develops. However, there appear to be more questions than answers about that ability development. For example, it is often not clear what distinguishes a learner at one ability level from a learner at another level and how a learner goes from one level to another. While we can describe learners at different ability levels, it is unclear what it is that the learner needs to do to get from one ability level to the next, or how much better the learner needs to perform to get to the next proficiency level.

As Bachman (1990) emphasized, communicative language ability consists of both knowledge and the capacity to implement it in different contexts of use. Therefore in the interactionist validation of a test, a person's performance on a test is taken to indicate an underlying trait characteristic of that person, and at the same performance is also taken to indicate the influence of the context in which the performance occurs. Therefore the interactionist validation allows to infer from test performance something about both content-specific behaviors and person-specific, context- independent traits.

According to Mislevy and Yin (2011), conceptions of capabilities and competencies shape the assessment argument. A perspective on language learning and language use suggests the kinds of inferences we should draw about examinees, from what kind of evidence, in what kind of assessment situations. A construct in an assessment is the assessor's inevitably simplified conception of examinee capabilities, chosen to suit a purpose of a given assessment. From an interactionist perspective, we see people's capabilities, contexts, and performances as intertwined in a constantly evolving process. The assessor's intentions determine what the

construct is, which in turn holds implications for design decisions in regard to tasks, evaluation procedures, and the measurement models used to aggregate evidence.

A significant issue in language assessment has been a lack of a theory of what abilities or components of abilities are thought to contribute to language development, or whose absence or underdevelopment may cause weaknesses in student abilities. This is in marked contrast with other areas of diagnosis, such as medicine, psychiatry, motor mechanics, or even first language acquisition issues (Alderson, 2007). Similarly, Bachman (2007) confirms that a persistent problem in language assessment has been that of understanding the roles of abilities and contexts, and the interaction between these, as they affect performance, or language assessment tasks.

Related to the difficulty of construct identification and specification, another issue in language assessment has been the assessment of academic language. For some researchers, this construct in itself appears unclear or artificial: since the only difference between conversational and academic language is vocabulary, with grammar, syntax, morphology, and phonetics being the same, is it legitimate to act like there is a big divide between the two? Can one really be isolated from another? Can one, but not the other, be taught? Can it be said that one has higher importance than the other? (Bailey & Butler, 2003).

Selection of Appropriate Assessment Tasks

According to Ross (2011), recently, efforts have been made to specify the relations among constructs, tasks devised to measure such constructs, and examinee performance characteristics. As performance assessment becomes increasingly more synonymous with task-

based assessment, specification of what an assessment task actually entails is subject to interpretive variability. While some authors identify tasks as a close approximation of what occurs in the real world, reproduced in an assessment context, others (Bachman, 2002) postulate that tasks designed without detailed specifications about the target domain can easily under-represent the construct these tasks are supposed to measure. Claims about language proficiency implied by performance on assessment tasks are thus crucially dependent on how thoroughly they sample and represent the constructs they are designed to measure and how they generalize to other contexts.

Performance assessments typically reveal one of the most problematic aspects of language testing. To the extent design specifications call for authenticity of how language is used in real life, the more potentially confounding factors are likely to be introduced. The more authentic the test is constructed to be, the more variance is introduced into the task, including, for example, the characteristics of the interlocutor (accent, speed etc.) and topics of the conversation. Additionally, it may be hard to ensure consistent rating of the examinee in an authentic situation.

Relationship between Language Domains

Alderson (2005) mentions that while language dimensions implicated in language learning can be easily outlined and named, the evidence rather than speculation about their relevance is relatively scarce. Similarly, standards of language achievement, which supposedly define the levels of attainment expected of language learners are often vague, ill-defined, lack empirical base, and bear little relation to theories of second language acquisition (p.21). Therefore, it is far from clear exactly what changes as learners develop and therefore what

diagnosis of second language development (or lack of it) should be based on or how diagnostic tests may be validated.

Some argue that English proficiency tests and the construct of language proficiency may be driven by the standards imposed by the NCLB, rather than by the linguistic theory. According to NCLB (2001), states must measure language proficiency and show progress; assess all ELL students; independently measure the four skill domains of reading, writing, speaking, and listening; report a separate measure for reading comprehension; assess proficiency in academic language and in the language of social interaction; and align the assessments with their state English Language Development (ELD) standards. Therefore in effect, NCLB imposes the four-domain structure on all compliant English language proficiency tests (Wolf et al., 2008). This mandate was interpreted such that states must report scores in each of the domains and an overall proficiency level. The test also reports an overall proficiency level which combines the four tests based on the rules adopted by the state. States were given the discretion to make the decision about how scores should be combined to create the overall proficiency level. There may not be a clear agreement on whether the present test structure is truly driven by the theory behind second language acquisition (Alderson, 2007).

In addition, there is some inconsistency in the determination of factor structure of such a well-researched language proficiency test as the Test of English as a Foreign Language (TOEFL). For example, some researchers (Swinton & Powers, 1980) found only one factor that is measured by the test. Similar results were obtained by Hale, Stansfield, Rock, Hicks, Butler, and Oiler (1988) using both factor analysis and IRT-based methods for assessing dimensionality, by Dunbar (1982) and Hale, Rock, and Jirele (1989), who used confirmatory factor analysis approaches, and by Boldt (1988), who used latent structure analysis. A study by McKinley and

Way (1992), who used MIRT analysis, suggested a consistent two-factor structure of the TOEFL test. Oltman, Stricker, and Barrows (1988) used a three-way multidimensional scaling approach to examine the effects of native language and English proficiency on the structure of the TOEFL test and found evidence of four possible factors.

Nevertheless, the structure of the test as consisting of domain skills in the four domains – listening, speaking, reading, and writing - appears to be rather common, with the understanding that the skills framework should be treated both individually and in an integrated manner (Jamieson et al., 2000). Sawaki, Stricker, and Oranje (2009) concluded that the current consensus in the field of language testing is that second language ability is multicomponential, with a general factor as well as smaller group factors. Few studies with the confirmation of factorial structures have been done on ELP tests for K-12 students; the majority of studies have been done on TOEFL, with the older student population in mind. However, a number of those studies confirm the appropriateness of subscore reporting by four domains both from the data analysis? and theoretical perspectives. For example, in a study by Sawaki et al. (2009) CFA analysis was performed testing five different models, a bifactor model, correlated four-factor model, single factor model, correlated two-factor model, and higher-order factor model. They reported a second-order general English as a second language factor and four first order factors for reading, listening, speaking and writing. A study by Shin (2005) confirmed a factor structure with a higher order overall proficiency factor and four lower-order factors for the domains. Correlational studies by Liu and Costanzo (2013) and Bozorgian (2012) indicate that the four domains measure distinct but related constructs. Powers (2013) states that while it is not always necessary to test language proficiency in all four domains (reading, writing, speaking, listening) if one is only trying to determine proficiency needs for a specific purpose, such as job

performance requirements, it is advisable to test all four domains, as the determination of proficiency gained from that assessment provides the best information for teaching and learning.

Differing Views of Professionals on Language Testing

Chapelle (2011) raises the issue of conflict between language proficiency as viewed by applied linguists and measurement professionals. While in applied linguistics, multiple views on the nature of language and on the construct of language proficiency tend to coexist, measurement professionals tend to have a constructivist view that allows predetermined language capacities of interest to be modeled in a way that is effective for specific purposes. Based on that constructivist position, Mislevy (2009) pointed out that constructs do not have to be psychologically true or accurate in order to be useful. They simply need to capture relevant response consistencies. Since models, or constructs, are always simplifications of complex phenomena, the question is not whether they are accurate, but rather how wrong they have to be in order to not be useful. This pragmatism is productive for language assessment; however, it can face criticism from linguists based on whether it reflects the true nature of language. Essentially, the opponents in such arguments dispute what is theoretically correct vs. what is pragmatically useful.

The argument between linguistics and measurement professionals is related to the construct definition underlying the test score interpretation. McNamara and Roever (2006) observe that traditionally, psychometrics prescribed "the rules of measurement, and language was virtually poured into these pre-existing forms" (p.27). However, as the language test use evolved, a better way of construct definition became necessary. Currently, Messick's (1981) framework for defining constructs is often used: constructs as samples of behavior in particular

contexts; constructs as signs of underlying traits; and constructs as signs of traits as they are sampled in particular contexts.

Language Assessments as Instruments of State Policy towards Non-native Speakers

Like all educational assessment, language testing is fundamentally concerned with social and distributive justice (Rawls, 2001). Shohamy (2006) listed three ways in which language policy objectives are achieved by language tests. Tests are instrumental in determining the prestige and status of languages (and thus maintaining the power of speakers of prestigious language varieties); standardizing and perpetuating language correctness (and thus maintaining the subordinate status of speakers of non-standard varieties); and suppressing language diversity (in favor of speakers of the prestigious standard variety). Many countries position their language policy as public monolingual despite their own multilingualism and therefore require a standard variety of the dominant or official language for immigration or citizenship. These procedures are set in place in order to ensure civic nationalism, social cohesion and national harmony, once again fusing language proficiency, citizenship, and identity. The role of a country's language policy is critical in determining its language assessment policy and practice towards potential immigrants, immigrants, and potential citizens (Kunnan, 2011).

The issues above have a direct impact on the validity and reliability of language proficiency assessment. These issues involve multiple professionals and require extensive multidisciplinary research to be resolved. They are critical to consider while examining a language proficiency assessment and subscore reporting for such an assessment.

Subscore Estimation and Reliability Evaluation Procedures

Procedures for subscore reporting may be classified for practical purposes into those that do or do not use information augmentation methods, and those that use CTT or IRT framework for parameter estimation. There is an overlap between the two classifications. For example, methods based on the CTT framework can use both non-augmented subscore reporting methods, such as raw score number correct subscore calculation, and augmented methods, such as Wainer's (2001) augmentation method. Similarly, methods of subscore reporting based on the IRT framework may not use data augmentation methods, such as raw IRT subscore ability calculation, or they may use augmentation methods, such as Bayesian IRT subscore reporting. In this section, we will review how the two common score scaling frameworks (CTT and IRT) approach subscore reporting, and how augmented vs. non-augmented subscore reporting methods work.

Subscore Estimation within the CTT Framework

Subscores may be computed under the paradigm of Classical Test Theory (CTT) or Item Response Theory (IRT), the two most prevalent models for test score scaling (Skorupski, 2008). These two models differ significantly in how they conceptualize reliability. In the CTT framework, the concept of reliability has to do with the consistency of test results: would we get the same score or a different one if we were to retake the same test (Thissen, 2000)? In CTT, the observed score of an examinee on a test (X) is composed of a true score (T), which reflects the test taker's true ability in the construct being measured, and an error score (E), which reflects the contribution from factors other than his/her true ability:

$$X=T+E$$

True score is defined as the average score over two parallel forms of the test. Parallel forms are defined as tests with observed scores p_1 and measure the same content and for which examinees have the same true scores (p_1' and p_2'), true score variance, and true error variance.

The following common approaches to subscore estimation can be used within the CTT framework: Kelley's regressed score method (1927), Wainer et al.'s (2001) multivariate empirical Bayes estimation method, and objective performance index scoring (OPI; Yen, 1987).

Kelley's Regression Method

One of the early regression-based methods of score estimation was Kelley's regression method. It is based on weighting the observed subscores based on the group mean. Using the CTT notation, it can be written as follows:

$$\hat{\tau} = \rho\chi + (1-\rho)\,\mu,$$

where $\hat{\tau}$ represents an estimate of true score (τ), χ is the observed score, μ is the group mean, and ρ is the reliability of the test. This method aims to improve the estimate of true score through the shrinkage in the observed score toward the group mean by an equal amount of reliability. When the test is very reliable, the impact of the observed score becomes very dominant on the estimate of the true score. However, when the test is not highly reliable, the estimate of the true score shrinks toward the group mean to remove the unreliable part of the observed score. Therefore, Kelley's regressed score method improves the precision of test scores by using the group mean as collateral information. Thus, the extent to which the collateral information (in this case, the group's mean) influences the estimate of the subscore is a function of the unreliability of the subscore measure. Considering a test with several subtests, Kelley's regressed score method can be applied to the subscores simultaneously (Skorupski, 2008).

Yen's OPI Method

Yen's OPI (1987) is a procedure that combines performance on a particular subtest with information from the examinee's overall test performance. This goal is accomplished by combining an empirical Bayes approach with an IRT domain score estimation method (Bock, Thissen, and Zimowski, 1997). Yen (1987) deals only with multiple-choice (MC) items, and Yen et al. (1997) extends the method to items of mixed types. For a multiple choice item test, the OPI T for subtest j, T_j, can be expressed by the following equation:

$$T_j = w_j \frac{x_j}{n_j} + \left(1 - w_j\right) \hat{T}_j ,$$

where $\frac{x_j}{n_j}$ is the observed proportion correct score for subtest j, $\hat{T}_j$ is the estimated proportion correct score for subtest j given $\hat{\theta}$ based on the total test using the IRT domain score method

$$\hat{T}_j = \frac{1}{n_j} \sum_{i=1}^{n_j} P_{ij}\left(\hat{\theta}\right),$$

and w_j is the relative weight given to the observed proportion correct score. From a Bayesian perspective, $\hat{T}_j$ is the prior containing information from the total test, which is usually more reliable; $\frac{x_j}{n_j}$ is the data containing information only from observed scores of subtest j; w_j and $1 - w_j$ represent weights assigned to the data and the prior, respectively; T_j is the posterior mean resulting from combining the prior with the data. Both $\hat{T}_j$ and w_j are estimated from the data. While Yen's, Kelley's, and Wainer's methods are similar in that they augment the subscores with prior information, they obtain this prior information from different sources (Yen – from the total score as the prior; Kelley – from the group mean as the prior; and Wainer – as a vector of other subscores as a prior).

Reliability Estimation within CTT Framework

The reliability coefficient can be formulaically expressed as the correlation between the scores of two parallel forms, or by the ratio of true score variance to observed score variance in the population of persons from which examinees are assumed to be randomly sampled (Lord & Novick, 1968; Green et al., 1984):

$$\rho_{xx} = \frac{\sigma^2_{true}}{\sigma^2_{obs}}$$

where σ^2_{true} is the variance of the true score, and σ^2_{obs} is the variance of the observed score in the group.

Reliability is usually expressed by one number (between 0 and 1) that summarizes the reliability of the whole test. The more reliable the measure is, the closer the coefficient value is to 1. While the reliability coefficient is easy to understand and interpret, and is also not unit-specific, it has some drawbacks that are grounded in the limitations of CTT in general. For example, like other sample-dependent measurements, it is most useful when the examinee sample is similar to the examinee population for whom the test is being developed. To the extent that the sample differs in some unknown way from the population, the utility of the item statistics may be reduced (Hambleton & Jones, 1993). Reliability coefficients are also dependent on the heterogeneity of the sample; the same test taken by a heterogeneous sample will have a higher value of ρ than when taken by a homogeneous sample (Thissen, 2000).

There are a number of ways to calculate reliability coefficient in CTT, such as, for example, Cronbach's alpha, Spearman- Brown prophecy formula, Kuder-Richardson 20 and Kuder-Richardson 21 (Crocker & Algina, 1986), stratified alpha (Cronbach, Schonemann, & McKie, 1965), and McDonald's omega (Gu, Little, & Kingston, 2013), to name a few. Despite some limitations, such as, for example, not being able to meet the assumption of essential tau-

equivalence (i.e., when all items have about equal true-score information), Cronbach's alpha remains a popular measure of reliability. Based on the definition of true test score in CTT,

$$\rho_{xx} = \frac{\sigma^2_{true}}{\sigma^2_{obs}} = \frac{\sigma^2_{t1+t2+\ldots.+tn}}{\sigma^2_{x1+x2+\ldots+xn}},$$

and Cronbach's coefficient alpha is defined as

$$\alpha = \frac{k}{(k-1)}\left(1-\frac{\sum \sigma^2_{x_i}}{\sigma^2_x}\right),$$

where k = number of items, $\sigma^2_{x_i}$ is the item variance, *and* σ^2_x is the person variance.

As Crocker and Algina (1986) note, coefficient alpha can be used as an index of internal consistency; it can also be considered as the lower bound to a theoretical reliability coefficient, known as the coefficient of precision. It is not a direct estimate of reliability, but rather an estimate of the lower bound of that coefficient. If we obtain a specific coefficient alpha, it is not possible to tell whether the coefficient of precision is actually higher than that estimate, or how much higher it may be.

Measurement of SEM in CTT

Within the framework of CTT, the reliability coefficient is inversely related to the standard error of measurement (Crocker & Algina, 1986). Since

$\sigma^2_{obs} = \sigma^2_{true} + \sigma^2_{error}$,

σ^2_{error} is essentially the average of all individual examinees' errors around their scores squared.

Since the reliability estimates provide information on a specific set of test scores, it cannot be used directly to interpret the effect of measurement on test scores for individual test takers, the

standard error of measurement, which is defined as the standard deviation of the error scores, is introduced for this purpose (Harvill, 1991). The standard error of measurement can be used to calculate confidence intervals for observed scores or true scores (He, 2009).

Since

$\sigma^2_{true} = \sigma^2_{obs} - \sigma^2_{error} = \sigma^2_{obs}$ – average SEM^2,

$SEM = \sigma_{obs}\sqrt{(1 - \rho_{xx})}$.

SEM is an index of the size of the errors, an index of the unreliability of a test, and is equal to the standard deviation of the error scores. If the error variance is large, it implies that the measurements were not precise and therefore the forms were unreliable (Lord & Novick, 1968). According to Green et al. (1984), SEM is a better measure of test reliability than the reliability index, even despite the fact that it is similar for all examinees. The main disadvantage of the SEM is that it is scale specific, and therefore SEMs for tests on different scales cannot be directly compared. The error variance, however, can be used to compare the different sources of error. Equation (3) is most often employed in estimating the error variance because it can sum variances attributed to all sources of error (Feldt & Brennan, 1989).

Based on how SEM is calculated, it is clear that since it is an average of individual error variances, it should represent the same value for all examinees. However, it was noticed by earlier educational researchers (e.g. Holzinger, 1921, as quoted by Gulliksen, 1950, and Mollenkopf, 1949) that this may not be the case in real-life applications of CTT; later on this assumption has been questioned by Feldt and Brennan (1989), Thissen (2000), and Haertel (2006). Specifically, it was noticed that CTT-based SEM tend to be smaller at the extremes of the scores, and larger in the middle range of the scores. The explanations of this phenomenon

were that, for example, the response patterns of those examinees who guess a lot (presumably due to lack of ability) are less consistent than of those examinees whose ability is high (Jarjoura, 1986); or that there is a general consistently consistent or consistently inconsistent patters of examinee responses that result in these SEM variations (Berdie, 1969). Green et al. (1984) suggests that the students who score very high or very low are likely to perform very consistently; the score is very repeatable, so it is very reliable. The CTT-based SEM is higher in the middle because scores are less repeatable there. Similarly, Hambleton & Swaminathan (1985) propose that examinees of higher ability perform a given task more consistently than medium-performing examinees, therefore their variance of error measurement will be smaller than that of medium-performing examinees.

Measurement of CSEM in CTT

It is clear that it may be very helpful to have the information about SEM at specific score levels – conditional SEM, or CSEM – to evaluate the precision of the test at those score levels. Recently, interest in CSEM measurement has grown because of recommendations that test publishers report them (AERA, 2014) and because of advances in methods for estimating CSEM both for raw scores and for the scale scores typically reported to examinees (Brennan & Lee, 1999; Feldt, 1984; Feldt, Steffen, & Gupta, 1985; Kolen, Zeng, & Hanson, 1996; Lee, Brennan, & Kolen, 2000; Lord, 1984; Woodruff, 1990).

One issue with measuring CSEM in CTT is that to even conceive of the idea that CSEM is not equal to average SEM and therefore is not the same across all values of ability contradicts the assumptions of CTT. CTT is a weak theory in that it does not make strong assumptions, and specifically, it does not make any assumptions about the distribution of ability. Additionally,

item characteristics are not tied in with ability, like in IRT. Therefore, one opinion may be that CSEM concept itself does not exist in CTT, or is not comparable with that of CSEM in IRT. Another opinion is that CTT and IRT measures of item and person parameters and reliability can be compared if some assumptions are relaxed (cf. Dimitrov, 2002, 2003; Kim & Feldt, 2010; Culpepper, 2013).

Feldt, Steffen, and Gupta (1985) describe five possible approaches to estimating CSEM in the CTT framework. Among the first methods proposed to estimate SEM for a specific score level, once it as recognized that CSEM does vary across score levels, was Thorndike's (1951) method. This method is based on the idea that the correlation between two parallel test forms' variance is 0. Therefore,

$\sigma_{Error} = \sqrt{\sigma_{e1}^2 + \sigma_{e2}^2}$.

Since $X_1 - X_2 = (T_1 - T_2) + (E_1 - E_2)$, and for parallel test forms $(T_1 - T_2) = 0$, then

$\sigma_{(x1-x2)} = \sigma_{(e1-e2)} = \sqrt{\sigma_{e1}^2 + \sigma_{e2}^2} = \sigma_{Error}$.

Therefore examinees can be grouped at specific score intervals, and then the standard deviation of half-test differences for a given subgroup can be an estimate of CSEM at the specific score level. One limitation of this method is that at some score levels, there could be many more examinees than at other levels (e.g. at the ends of the score scales), which will skew the calculation of standard deviations of score differences.

The second method, proposed by Mollenkopf (1949), is referred to as the polynomial method. It is based on the same idea of parallel forms, but uses least squares regression method to "predict" the squared difference between the half scores from the total score for each individual. Suppose that Y is the squared difference between half-scores:

$Y = (X_1 - X_2)^2$

A quadratic, cubic, or fourth degree polynomial can be used to predict this value. For example, a fourth degree polynomial model to predict $\hat{Y}$ will be

$\hat{Y} = a_0 + a_1(X) + a_1(X^2) + a_1(X^3) + a_1(X^4)$.

$\hat{Y}$ may be interpreted as an estimate of the average of Y for the group of students who have a score of X. Since with parallel tests, the average squared difference between the two scores is the variance of differences, it is the variance of errors of measurement for the full test. Therefore $\sqrt{\hat{Y}}$ for a given value of X is an estimate of SEM at score point X. This technique is more effective than grouping examinees in score intervals, as it smoothes and stabilizes the estimates.

A third estimation technique is based on Lord's (1955) binomial error model. Suppose there is a universe of test items from which a number of items, k, constitutes a test form. Another independent set of k items constitutes a second form, and so on. A given individual i is thought to be able to answer a certain proportion, φ_i, of the entire population of items. The fundamental notion of the standard error of measurement is that of the standard deviation of scores for a given examinee on many parallel test forms. Thus, the concept of σ_E is directly comparable to the statistical concept of a standard error of a frequency. That is, the standard error of a frequency determined from a sample of size k is $\sqrt{[k(\varphi_i)(1-\varphi_i)]}$ for a person i who is able to answer the φ proportion of the items on form k. Since the parameter φ_i is unknown, Lord (1955) proposed the use of the observed proportion correct as an estimate of φ_i. He also recommended correction for the known bias in the variance determined from finite samples. The end result was the following formula as the estimate of the standard error of measurement at the score level X for the total test:

$$CSEm = \sqrt{[\frac{X(k-X)}{k-1}]}.$$

This formula, however, does not take into account the fact that forms may differ in content, item difficulty, and other characteristics, and therefore overestimates SEM at score value X.

Keats (1957) proposed an adjustment to this formula, in which the reliability of the form is taken into consideration:

$$CSEm = \sqrt{[\frac{X(k-X)}{k-1}]\ (\frac{1-r_{xx'}}{1-r_{21}})}.$$

Lord (1965) proposed a modification that takes into account the fact that the items on form k may have different characteristics and therefore need to be grouped together based on those characteristics:

$$CSEm = \sqrt{\sum_{h=1}^{c} \frac{X_{ih}(k_h - X_{ih})}{k_h - 1}},$$

where X_{ih} is the score of person i on the cluster of items corresponding to category h of the test specifications, c is the number of item categories, and k_h is the number of items in category h.

Another approach by Hoyt (1941) draws upon analysis of variance (ANOVA) methodology. The examinees by items score matrix for a test may be analyzed to obtain mean squares for examinees (MS_S), items (MS,), and interaction (MS_{s*i}), and the reliability of the test may be estimated from these mean squares. The error variance for a test of k items may be approximated by k(MSsx¡)' If individuals are grouped into intervals of total score, the standard error of measurement for the interval midpoint may be estimated by

$$CSEm = \sqrt{[k(MS_{s*i})]}.$$

Some methods of CTT-based CSEM calculation involve IRT-based methods. Feldt et al. (1985) concluded that these methods above are pretty comparable in their end results, with the largest

differences among the methods occurring in the CSEM values for the most extreme score intervals. The measurement error variance may be estimated for examinee i by a two-step process: (1) obtaining an estimate of the examinee's θ_i; and (2) evaluating the function

$$Se_i^2 = \sqrt{\{\sum_j^k [P_j(\theta_i)][1 - P_j(\theta_i)]}.$$

$P_j(\theta_i)$ is the value of the function for item j when evaluated at the ability level, θ_i, of subject i. To use this method, examinees are grouped according to their level of total score, the average value of S2,(i) is computed for the examinees in each interval, and the square root of the average is determined for the interval. An alternative approach is to derive a table of one-to-one correspondences between values of X and values of 0. Each specific pair (X_0, θ_0,) is derived by determining the value of θ , which satisfies the relationship

$$X_0 = \sum_j^k P_j(\theta_0).$$

The squared standard error at value $\theta_0$is then associated with the raw score value X_0.

To conclude, a number of methods have been devised to acknowledge the fact that CSEM may not be evenly distributed in CTT-scored tests, and to compensate for the lack of information regarding different SEM values at different score levels.

Subscore Estimation within the IRT Framework

General IRT Scaling Principles

In the non-augmented IRT-based scoring, either a unidimensional item calibration is performed using all items from a test, and then subscores for specific subscales are calculated using only those items that apply to each individual area, or separate calibrations are performed

within each content area, treating each subscale as its own test, and then subscores are reported for each calculated subscale (Skorupski, 2008). In either case, only the information about the individual items is taken into consideration to calculate the probability of a given response pattern (figure 2).

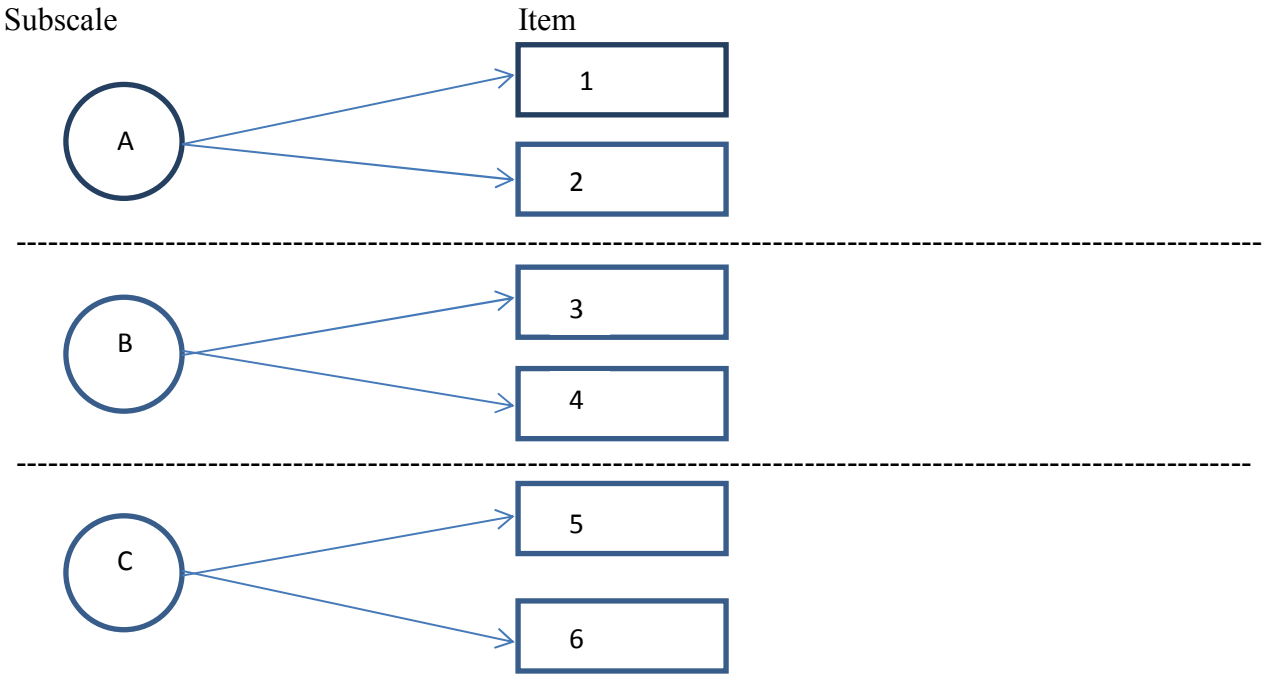

Figure 2. Test scoring with UIRT: subscale correlations are not considered.

Unlike CTT, IRT makes a clear connection between ability and item difficulty by putting them on the same scale. The relationship between ability and the probability of answering a dichotomously scored item correctly in the IRT framework is expressed as follows:

For a 1 PL model:

$$P_i(\theta) = \frac{e^{D\bar{a}\,(\theta - b_i)}}{1 + e^{D\bar{a}\,(\theta - b_i)}};$$

for a 2 PL model:

$$P_i(\theta) = \frac{e^{Da_i(\theta - b_i)}}{1 + e^{Da_i(\theta - b_i)}};$$

And for a 3 PL model:

$$P_i(\theta) = c_i + (1 - c_i)\frac{e^{Da_i(\theta - b_i)}}{1 + e^{Da_i(\theta - b_i)}},$$

where $P_i(\theta)$ is the probability of an examinee with ability level θ answers item I correctly;

b_i = the item difficulty parameter;

a_i = the item discrimination parameter;

$\bar{a}$ = the constant which is the common level of discrimination for all items;

D=1.7 (the scaling constant) (Hambleton & Swaminathan, 1985).

The relationship between item parameters and ability is visually represented in the item characteristic curve (ICC; figure 3). The difficulty parameter (*b*) is the point on the ability scale which corresponds to the point of inflection on the ICC. The higher the *b* value, the more difficult the item is. The 2 PL model adds the *a* parameter (discrimination) that is proportional to the slope of a line that is tangent to the ICC at point *b* on the ability scale. The steeper the slope of the tangent line, the more discriminating the item is. The 3 PL model adds the (pseudo)guessing parameter *c* that is the lower asymptote of ICC. Since even the lowest ability examinees have some probability of guessing the correct answer, this probability is reflected in the *c* parameter.

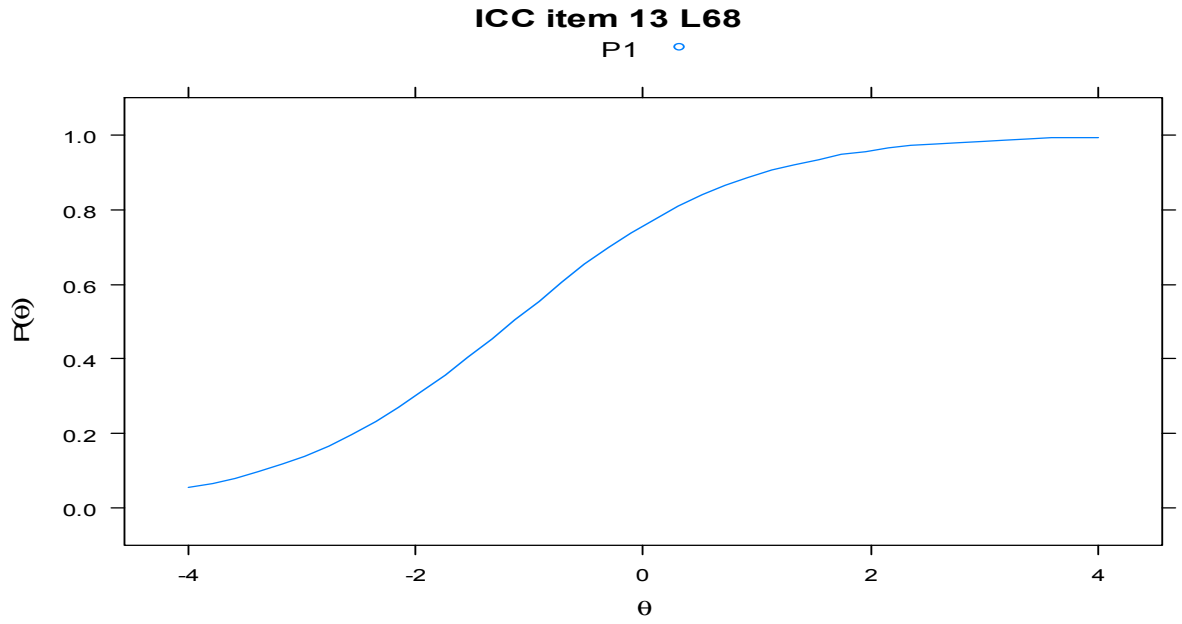

Figure 3. Item characteristic curve.

The test characteristic curve combines the characteristics of the items that contribute to the test (figure 4).

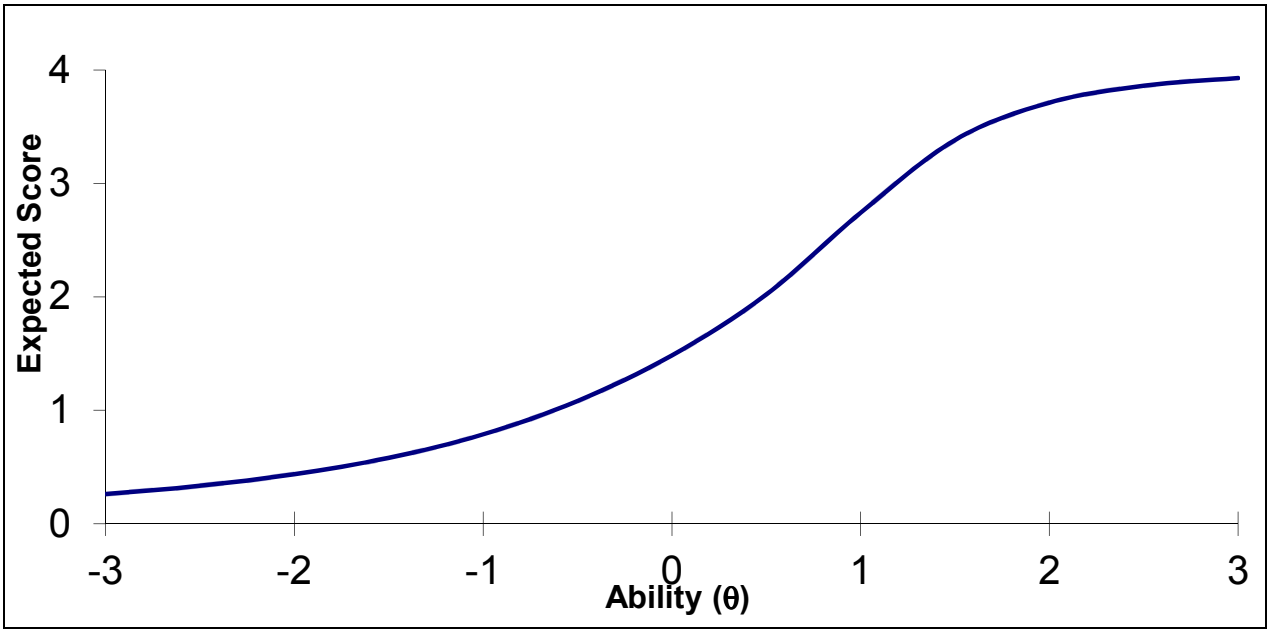

Figure 4. Test characteristic curve.

The test information function is a concept that indicates the utility of a particular test for evaluating different levels of examinee ability. As a test is composed of items, so is the test information function composed of item information functions (figures 5, 6; Skorupski, 2008).

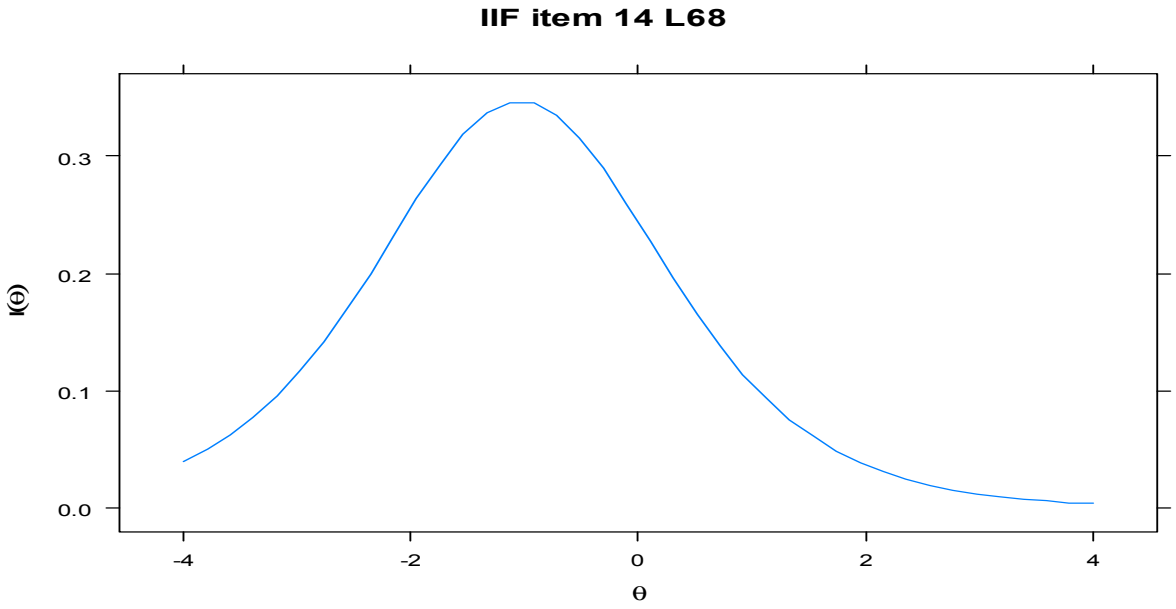

Figure 5. Item information function.

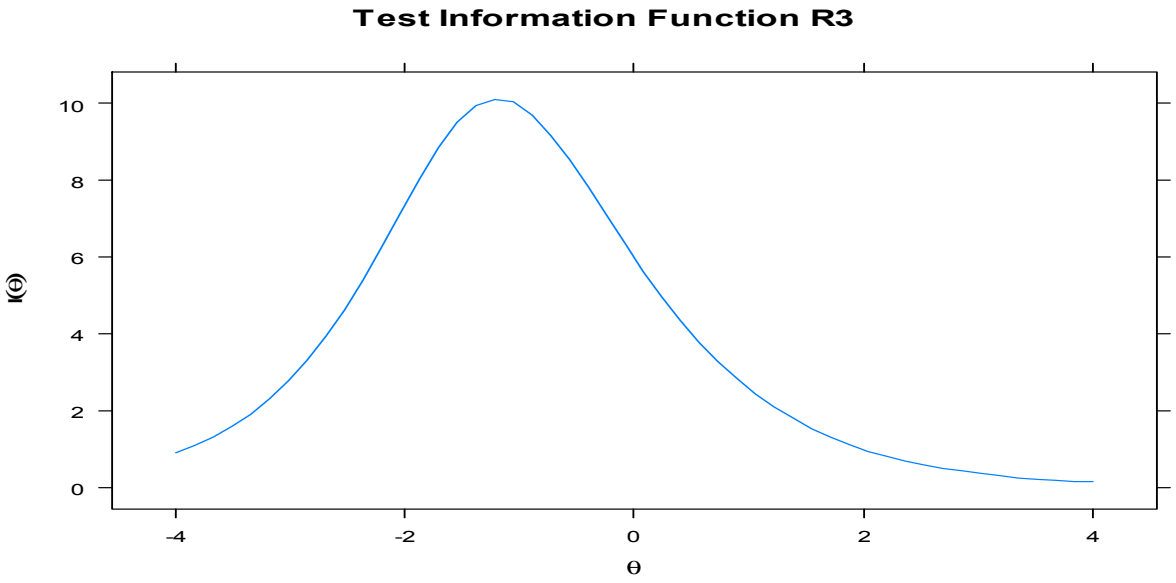

Figure 6. Test information function.

The information of a test is therefore the sum of individual item information functions at specific levels of θ:

$$I(\theta) = \sum_{i=1}^{n} \frac{P_i'^2}{PiQi},$$

Where P_i' is the derivative of the item response function with respect to θ; P_i is the likelihood of a correct response. Specifically, the formulas for the information function for 1, 2, and 3 PL logistic IRT models are as follows:

For 1 PL:

$I(\theta) = \sum_i D^2 P_i \ Q_i$;

for 2 PL:

$I(\theta) = \sum_i D^2 a_i^2 P_i Q_i$;

and for 3 PL:

$I(\theta) = \sum_i D^2 a_i^2 Q_i (P_i - c_i)^2 / (1 - c_i)^2 P_i$,

where $P_i(\theta)$ is the probability of an examinee with ability level θ answers item I correctly;

b_i = the item difficulty parameter;

a_i = the item discrimination parameter;

c = the guessing parameter;

D = 1.7 (the scaling constant) (Hambleton & Swaminathan, 1985).

From these formulas it becomes obvious that item discrimination and difficulty influence the information function. Information is highest in the items with the largest discrimination parameter and the difficulty parameter that equals the ability (in other words, where the examinee has a 0.5 probability of answering the item correctly).

IRT Models Including Polytomous vs. Dichotomous Items

While much of the data in educational measurement is scored dichotomously (e.g. 1=correct response; 0=incorrect response), in some situations the data may have more than two response categories. These models can be classified based on the number of item parameters they allow as Rasch or non-Rasch models. One of the examples of a Rasch ordered polytomous model is the partial credit model, in which partial credit is assigned for partially correct responses, thus allowing this additional information to help estimate the person's location on the ability continuum. For example, in a math item that requires a student solve a problem (4+5) x 3 = ?, the student may be able to do the addition part, but not the multiplication part. In an all-or-nothing situation, the student will get 0 points (no credit) unless he or she is able to perform both the addition and the multiplication steps correctly. In a partial credit situation, he or she will get 0 points for being unable to perform both addition and multiplication; 1 point for performing the addition only; and 2 points for performing both addition and multiplication correctly. Another example of a Rasch polytomous model is the rating scale model, in which responses to a series of items come from a series of ordered categories that are separated from one another by a series of ordered thresholds. For example, the person may be asked to rate how often they feel tired as part of answering a depression scale questionnaire: never, sometimes, often, or always.

Examples of non-Rasch models for ordered polytomous data that relax the assumption of equal item discrimination include the graded response model and the generalized partial credit model. The graded response model specifies the probability of a person responding with a category score x_j or higher vs. responding to lower category scores. In essence, the polytomous score has been turned into a series of cumulative comparisons (below a particular category vs. at or above this category, or getting a score of 2 vs. a score of 0 or 1) (De Ayala, 2009).

Since the generalized partial credit model was used to score the items, this model for polytomous items scoring will be reviewed in greater detail. Generalized partial credit model (GPCM) was developed by Muraki (1992) on the basis of the partial credit model (PCM) developed by Masters (1982). The PCM allows an individual to receive credit for answering an item that consists of several steps partially correctly: that is, by performing only some of the steps correctly, but not all of them. If there are two steps in the item, the person may perform the first step correctly, but not the second one. Thus, if the item was scored based on the correct performance of all steps (all or nothing), the person would get no credit. However, under partial credit model the person would get partial credit for performing one step correctly. The assumption is that if the first step is not performed correctly, the second step cannot be performed correctly, so the person cannot get credit for step two without accomplishing step one correctly. Note that the steps are not ordered by difficulty; that is, the first step may be harder than the subsequent steps. The scores for each step, or category, are called category scores and can be taken to indicate the number of successfully performed operations. Each transition point is conditionally defined and cannot be interpreted independently. Higher-category scores thus indicate a higher level of overall performance than do lower category scores. In the partial credit model the responses are decomposed into a series of ordered pairs of adjacent category scores and then successively applying a dichotomous model to each pair. Each adjacent response category is separated by a transition point that the person needs to pass to get into the next category. Each of those transition points is represented by a b value, which is the transition location parameter. The partial credit model is thus a Rasch model, in that all items are assumed to have the same discrimination parameter, but each item has several transition location parameters indicating successful transition from one category to the next. In other words, the

transition location is the point where the probability of responding to two adjacent categories is equal.

On an item that has two categories and 3 possible scores (0, 1, 2), the individual can get a score of 2 only if he passes from point 0 to point 1 and from point 1 to point 2. The probability of both events occurring is given by the adding the probability of two mutually exclusive events of passing from 0 to 1 occurring, and passing from 1 to 2 occurring. A general expression that incorporates the principles of PCM specifies that the conditional probability of an examinee with latent location θ obtaining a category score of x_j is

$$p\left(x_j \mid \theta, b_{jh}\right) = \frac{\exp[\sum_{h=0}^{x_j}(\theta - b_{jh})]}{\sum_{k=0}^{m_j}\exp[\sum_{h=0}^{k}(\theta - b_{jh})]},$$

where b_{jh} is the transition location parameter for item j; it is sometimes referred to as the step difficulty parameter, or step parameter. b_{jh} reflects the relative difficulty in endorsing category h over category (h-1). The examinee's responses are categorized into m_j+1 scores, and m_j is the number of operations the examinee needs to perform to correctly answer item j.

Muraki's 1992 GPCM relaxes the assumption that all the items on the test have the same discrimination parameter. Therefore, instead of using a dichotomous Rasch model, Muraki proposed a 2PL IRT model to express the probability of selecting a particular response category over the previous one. The transition locations in the previous equation can be decomposed into an item difficulty component and a threshold parameter as follows:

$$b_{jk} = b_j - \tau_k.$$

By substituting this expression into the previous equation, we will get

$$p\left(x_{jk} \mid \theta, a_j, b_j, \tau\right) = \frac{\exp[\sum_{h=0}^{k} a_j(\theta - b_j + \tau_{jh})]}{\sum_{c=1}^{m_j} \exp[\sum_{h=1}^{c} a_j(\theta - b_j + \tau_{jh})]},$$

where m is the number of response categories, k=1, …, m, and $\tau_1 = 0$. The discrimination parameter a_j indicates the degree to which categorical responses vary among items as θ changes. τ_h may be interpreted as the relative difficulty of step h in comparing other steps within an item; this difficulty may also be interpreted as the difficulty of endorsing a particular category (De Ayala, 2009). Similar to the PCM, the categories do not have to be in ascending order, as the initial categories may be more difficult to endorse than the subsequent ones (De Ayala, 2009).

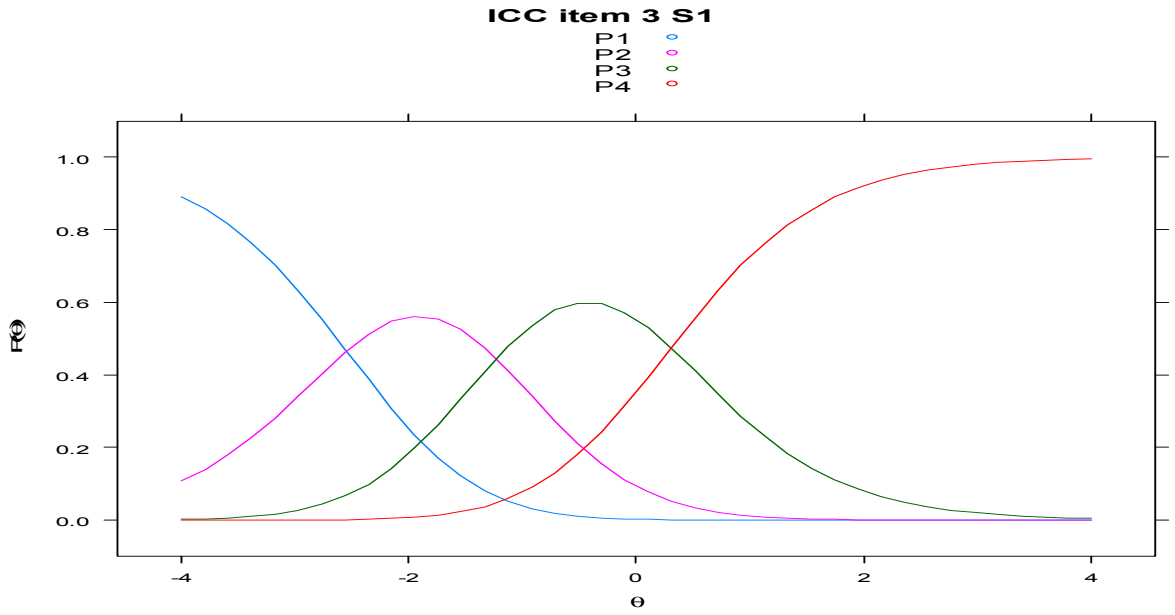

Figure 7. Category response functions for an item with four score categories (GPCM model).

Reliability Estimation within the IRT Framework

The concept of reliability in IRT has to do with correctly estimating the examinee's location along the ability continuum, rather than with expecting to obtain a consistent score after

administering a number of parallel forms to the examinee. CTT measures of reliability are sample-specific; in addition, error is estimated for a group, rather for individuals at specific score values. The item response analog of test score reliability and the SEM is the test information function (TIF). The use of TIF as a measure of accuracy of estimation is more appealing, because it is not sample-dependent, and it provides an estimate of the error of measurement at each ability level (Hambleton & Swaminathan, 1985).

Estimation of SEM in IRT

The reciprocal of the information function evaluated at ability θ is the asymptotic variance of the maximum likelihood estimator $\hat{\theta}$. This can be expressed as

$V(\hat{\theta} \mid \theta) = \frac{1}{I(\theta)}$,

with V denoting variance.

Consequently SEM = $\sqrt{V} = \frac{1}{\sqrt{I(\theta)}}$.

Therefore the information function is inversely proportional to the square of the width of the asymptotic confidence interval for θ. The larger the value of the information function, the smaller the width of the confidence interval, and in turn, the more precise the measurement of ability will be (figure 8). Since the information function is the measure of ability, it will have different values at different ability levels; hence, the precision of measurement can be evaluated at a specific ability level (Hambleton & Swaminathan, 1985). Since the test information is a function of person ability, the standard error of person ability is also a function of ability (He, 2009).

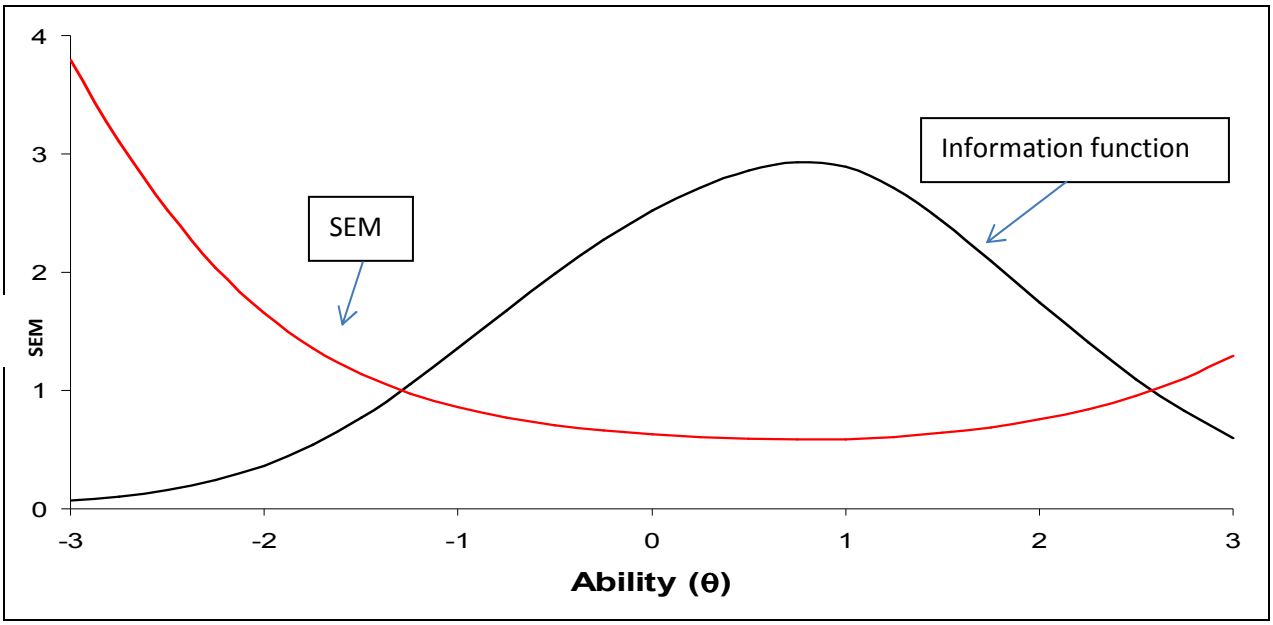

Figure 8. Information and SEM in IRT.

CTT and IRT Reliability Comparison

Since the formulas for IRT test-based reliability are not based on the same assumptions as CTT formulas for reliability, they do not lead to the same reliability coefficients (Kim & Feldt, 2010). As compared to the CTT calculation of SEM, it is apparent that the IRT SEM calculation inherently provides the error estimate at a specific ability level and is thus considered a conditional standard error of measurement. It is noted that, opposite to the CTT CSEM, the IRT that IRT ability estimates have smaller SEM in the middle range of scores can be explained by the fact that there are more people in the middle range of ability, which allows for a better estimate and discrimination of ability (Hambleton & Swaminathan, 1985). Additionally, items that are too hard or too easy for the examinees to whom they are administered (the difficulty at the two extremes of the continuum) are less likely to be informative than the items that are "just right" – the items that are of medium difficulty and are in the middle of the continuum.

Some researchers (c.f. Samejima, 1977; Kim & Feldt, 2010), stress the incompatibility of CTT and IRT frameworks, pointing out that the CTT-based definition of reliability has little relevance for measurement based on IRT, where the error variance is expressed as a function of ability. This difference in framework assumptions makes it challenging to compare reliability and precision estimates for scores derived within CTT and IRT frameworks.

At the same time, a number of researchers attempted to reconcile these differences between reliability and precision measurement in CTT and IRT (e.g. Lord, 1980; Green et al., 1984; Dimitrov, 2002, 2003; Wang, Cheng, and Chen, 2004). By following the same approach used for test reliability in CTT, Wang et al. (2004) described a method for obtaining IRT-based reliability. First, the test information is averaged over the θ level to obtain $\bar{T}$. The average test information is the average degree of measurement precision that the test or subtest provides for the sampled persons. Based on this fact, the IRT-based test reliability, which is also called the composite test reliability, can be defined as:

$$\rho_{IRT} = 1 - \frac{\bar{T}^{-1}}{\sigma_\theta^2},$$

where σ_θ^2 is the variance of the θ distribution. This IRT reliability is also known as marginal reliability. To simplify the computation of this reliability, Mislevy et al. (1992) suggested a simpler solution when MML estimation is used:

$$\rho_{MML} = \frac{\sigma_{EAP}^2}{\sigma_\theta^2},$$

where σ_{EAP}^2 is the variance of the EAP estimates. Wang et al. (2004) noted that the second formula of IRT test reliability is more practical in real data analysis. Some researchers developed methods of IRT reliability calculations that are based on CTT concepts. For example, Kim and Feldt (2010) described how to estimate an IRT-based reliability coefficient using the CTT

framework. From the perspective of nonlinear regression, Kim and Feldt (2010) argued that the same approach can be applied to the correlation of test score X with ability θ:

$$\rho_{xx\prime} = \frac{\sigma_t^2}{\sigma_x^2} = 1 - \frac{\sigma_e^2}{\sigma_x^2} \ ,$$

where σ_e^2 is the average test error variance over an ability distribution of the conditional error variances given θ.

Green et al. (1984) developed the measurement of marginal reliability estimation. The marginal reliability for $\hat{\theta}$ reflects the average accuracy across the ability continuum. However, this value accurately characterizes the precision of measurement across the continuum only when the total information function is uniformly distributed. The following formula expresses marginal reliability measurement:

$\rho\ (\theta) = \frac{\sigma_\theta^2 - \sigma_{em}^2}{\sigma_\theta^2}$.

As an alternative to an index for an entire metric, one may want to calculate conditional reliability at a given θ point. Green et al. (1984) calculate conditional reliability as

$\rho\ (\theta) = \frac{\sigma_\theta^2 - \sigma_e^2(\theta)}{\sigma_\theta^2}$,

where σ_θ^2 is the variance of observed ability, and $\sigma_e^2(\theta)$ is the expected variance error of the estimate. This would be the reliability if everyone were measured with the same precision as those persons with ability θ. Another method of estimating reliability is based on Linacre's (1997) formula:

$R_{IRT} = 1 - \frac{\sigma_{IRT,avg}^2}{\sigma_{0,IRT}^2}$,

where $\sigma^2_{IRT,avg}$ is the average of the person measure error variance, and $\sigma^2_{0,IRT}$ is the observed person measure variance (He, 2009).

Parameter Estimation in IRT

In the present assessment, due to the fact that item calibration has already been done, the item parameters are known. Therefore, what remains to do is estimate the person parameter (ability). Similar estimation procedures for ability estimation are used when item parameters are known to the ones that are used when item parameters are not known. Specifically, these procedures include maximum likelihood estimation (MLE); maximum a posteriori (MAP), and expected a posteriori (EAP). Joint maximum likelihood estimation (JMLE) and marginal maximum likelihood estimation (MMLE) procedures are used when both the item and person parameters are not known. Below we describe these procedures as applied to both item and person parameter estimation. We also illustrate how reliability and precision are estimated for these parameter estimation methods.

Maximum Likelihood Estimation (MLE)

Various strategies have been developed to solve the problem of estimating one set of parameters (e.g. the student ability parameter, θ) without the knowledge of another set of parameters (e.g., the item parameters). One of these approaches is maximum likelihood estimation (MLE). In this method, we may consider that we know the items' location (parameters), and then try to solve the problem "which θ values has the highest likelihood of producing a given response pattern?" The first step in this procedure is to calculate the probability of each item response in the pattern based on the IRT model, and the second step is to determine the probability of the whole pattern. Then these steps are performed repeatedly for a

range of θ values. This approach has a problem in that if a person has an extreme response pattern (all 1's or all 0's), his or her ability cannot be estimated using this approach. Similarly, if (on a short test) the responses to a given item are all 1's or all 0's, the item parameters cannot be estimated using MLE.

Joint Maximum Likelihood Estimation (JMLE)

Another approach maximizes the joint likelihood function of both persons and items in order to simultaneously estimate both person and item parameters. This strategy is known as joint maximum likelihood estimation (JMLE). The probability of having a correct score on an item given ability θ needs to be determined for a given person, and then that probability needs to be multiplied by the number of people who took the item. Therefore both item and person parameters are estimated simultaneously (De Ayala, 2009).

Marginal Maximum Likelihood Estimation (MMLE)

An alternative to this method of estimation is marginal maximum likelihood estimation (MMLE), in which person parameters are assumed to be known, and only item parameters are estimated. To be able to not have to directly estimate the person parameters, however, certain assumptions about ability distribution of ability parameter in the population have to be made. Now the probability of a given response pattern, instead of being conditioned on a given individual's θ, like in JMLE, is not conditioned on θ. Although a specific individual's θ is unknown, the probability of possible θs can be determined on the basis of the individual's responses, the item parameters, and the population distribution of ability. The ability distribution is divided into several quadrature points, and individuals are sampled from each quadrature node.

Ability Estimation in IRT

As mentioned previously, some procedures (e.g. JMLE) estimate the ability (person parameter) and item parameters simultaneously. For those procedures that estimate item parameters first, once that step has been accomplished, we can proceed to obtain the person parameters. This can be done using maximum likelihood (MLE); but if we want to avoid the inability of MLE to estimate parameters in case of extreme response patterns, we may want to consider two Bayesian estimation methods, maximum a posteriori (MAP) that uses the mode of the posterior distribution as the θ estimate, or expected a posteriori (EAP), the one that uses the mean of the posterior distribution as the θ estimate. All three methods of estimation (MLE, MAP, EAP) treat the item parameters' estimates as known and ignore their estimation errors when estimating $\hat{\theta}$. But, unlike MLE, both MAP and EAP can estimate parameters in the extreme response patterns. The similarity between MAP and EAP is that, like all Bayesian approaches, they incorporate information from other sources into the estimation of parameters. Both MAP and EAP, as Bayesian estimators, are regressed towards the means of prior distribution. For both methods, as Bayesian estimators, the standard error of estimation is the posterior standard deviation. The use of $PSD\left(\hat{\theta}\right)$ as the standard error is based on the fact that after 20 items the likelihood function and the posterior distribution are nearly identical, and the $PSD\left(\hat{\theta}\right)$ is pretty much interchangeable with the standard error (Bock & Mislevy, 1982; De Ayala, 2009).

The differences are as follows: the MAP approach is an iterative approach, like MLE, whereas EAP is non-iterative and is based on numerical quadrature methods like the ones used in MMLE. Because of this non-iterative, and thus more efficient, nature, EAP is potentially faster than MLE or MAP in estimating the person location. Second, EAP uses a discrete prior distribution, whereas MAP uses a continuous prior distribution. Third, whereas MAP $\hat{\theta}$s exist for

all response patterns, they suffer from greater regression towards the prior's mean than do EAP estimates (Bock & Mislevy, 1982). Fourth, the average squared error for EAP estimates over the population is less than that for MAP and MLE person location estimates (Bock & Mislevy, 1982). Fifth, the mathematical computations of person parameter is simpler in EAP. EAP estimate has been cited as favored over the MLE and MAP estimate, and is considered the method of choice by a number of authors (Wang et al., 2004; Mislevy & Stocking, 1989). Bock and Mislevy (1982) show that the EAP method produces reasonably accurate estimates of the person parameters. The accuracy of the EAP sample standard errors has also been studied (DeAyala, Schafer, & Sava-Bolesta, 1995).

The Bayesian estimation procedures utilize the likelihood function, or corresponding log-likelihood expression, but incorporate a prior density into the estimation to arrive at a posterior distribution of ability:

$f(\theta\,|X) \propto L\,(X|\theta) f(\theta)$, and

$\ln[f(\,(\theta\,|X)] \propto \ln[L\,(X|\theta)] + \ln[f(\theta)]$,

where $f(\theta)$ is the prior density of ability, which can be uniquely determined for each examinee or common across all examinees, $L\,(X|\theta)$ is the likelihood of observing a score pattern given the ability level, and $f(\theta\,|X)$ is the posterior distribution of ability. Inferences regarding examinee ability (e.g., expectation and variance) are then made based on the posterior distribution. The EAP estimate is determined by calculating the mean of the posterior distribution, while the MAP estimate is determined by finding the maximum value, or mode, of the posterior distribution (Lord, 1980; Hambleton & Swaminathan, 1985).

The EAP estimate of an individual's θ after administering L items is

$$\hat{\theta}_i = \frac{\sum_{r=1}^{R} X_r \, L\,(X_r) A\,(X_r)}{\sum_{r=1}^{R} L\,(X_r) A\,(X_r)},$$

Then the posterior standard deviation, which is the standard error of measurement for EAP, is

$$PSD\left(\hat{\theta}\right) = \sqrt{\frac{\sum_{r=1}^{R} (X_r - \hat{\theta}_i\,)^2 \, L\,(X_r) A\,(X_r)}{\sum_{r=1}^{R} L\,(X_r) A\,(X_r)}},$$

where R is the number of quadrature points; $X_r$is the midpoint of a quadrature interval known as the quadrature node or point and has an associated weight $A\,(X_r)$ which reflects the height of the function g ($\theta \mid \vartheta$) around X_r, and $L\,(X_r)$ is the likelihood function at X_r given a certain response pattern x=x_1 , …, x_L and a particular IRT model, such as a 1 PL, 2 PL, or a 3 PL model. In addition, Bock and Mislevy (1982) suggested a reliability coefficient for EAP location estimate,

$\rho = 1 - PSD\left(\hat{\theta}\right)^2$,
which is based on the assumption that the latent variable is normally distributed in the population with the mean of 0 and variance of 1.

Subscore Estimation within MIRT Framework

Multidimensional IRT can be used for subscore reporting and for increasing the reliability of subscores. The principle behind the MIRT approach is that each subscore represents a distinct trait; although a single total test score will ultimately be used for any decision making, there is an implicit assumption that the test is actually a mixture or related multidimensional traits (Thissen & Edwards, 2005; Ackerman, 1994; Luecht, 2003). As such, the MIRT approach to calculating subscores, like the other augmented approaches, also takes advantage of shared information across subscores to improve the reliability of these estimates (Skorupski, 2008). UIRT subscore reporting does not take into consideration the relationships between subscales;

MIRT approach allows one to benefit from the additional information about the person's abilities on all subscales of the test.

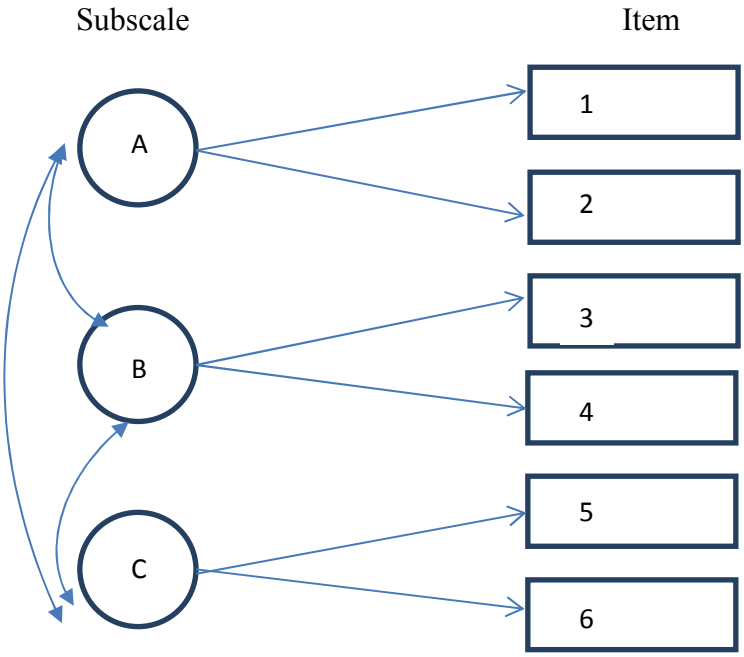

Figure 9. Test scoring under MIRT: subscales are assumed to be correlated.

The major advantage of MIRT is that using this model, subscores and the total score can be calculated in one procedure, and in addition multiple skills with complex interactions of persons and test items can be measured (Reckase, 2009). The major disadvantages currently are considered the complexity of estimation and model fit determination, and length of computation. However, the MIRT model can be used to better understand the traits being measured in a multidimensional context by providing an estimate of each latent trait, as well as an item's discrimination for each of these, and the overall difficulty of the item (Finch, 2011).

MIRT is generally viewed as a more complex, more computationally demanding, and less straightforward procedure of subscore estimation (Skorupski, 2008) than other subscore

estimation methods. It requires the knowledge of such procedures as making choices between subtly different MIRT models, choosing among estimators, performing confirmatory factor analyses for the number of dimensions fit, sorting out rotational indeterminacies (Luecht, 2003), and at times facing non-convergence of estimation solution (Skorupski, 2008), that may not be within the day-to-day scope of practice of a psychometrician working in the educational testing industry (Haberman & Sinharay, 2010). In addition, the results of a MIRT analyses are not always easily comparable with unidimensional methods, except in terms of model fit (Luecht, 2003).

Standard item response theory (IRT) models are based on the assumption that the latent trait, or ability, being measured is unidimensional in nature; that is, the items of interest are all associated with only one latent construct (Finch, 2011). However, few, if any, tests are perfectly unidimensional (Green et al., 1984). Indeed, a number of previous studies have shown that the unidimensionality assumption is often violated in real-world contexts (Ackerman, 1994; Nandakumar, 1994; Reckase, 1985), and the number of dimensions is underestimated (Reckase & Hirsh, 1991), which may lead to increased errors of measurement and the possibility of making incorrect inferences about a student's ability (Walker & Beretvas, 2003).

To address this issue, multidimensional IRT (MIRT) models for linking multiple latent traits with an item response were developed (e.g., Reckase, 1985). This MIRT model takes the same general form as the unidimensional 3PL model, reflecting the link between the latent ability and the probability of a correct item response through an item's discrimination, difficulty, and guessing parameters.

The UIRT measures of item difficulty and discrimination are directly related to the characteristics of the item characteristic curve (ICC). The difficulty parameter indicates the value

of θ that corresponds to the point of steepest slope for the ICC. The discrimination parameter is related to the slope of the ICC where it is steepest. These two descriptive statistics for test items can be generalized to the MIRT case, but with some caveats. In MIRT, the slope of a surface is dependent on the direction of movement along the surface, so the point of steepest slope depends on the direction that is being considered. At each point in the θ -space, there is a direction that has the maximum slope from that point. If the entire θ -space is considered and the slopes in all directions at each point are evaluated, there is a maximum slope overall for the test item. The value of the maximum slope would be a useful summary of the capabilities of the test item for distinguishing between θ -points in the direction of greatest slope. One also needs to know the relationship between the *a*-parameter vector and the values of the slopes at a point in the θ -space to evaluate how well the item is able to differentiate between θ -points at different locations in the space using estimates of the item parameters. The parallel measure of item difficulty is the distance from the origin of the θ -space (i.e., the 0-vector) to the θ -point that is below the point of steepest slope for the surface. The sign associated with this distance indicates the relative position of the θ -point to the origin of the θ -space. Given the distance to the point and the directions from the axes, the values of the coordinates of the point on each of the axes can be recovered using the trigonometric relationship

$$\theta_{\vartheta} = \gamma \cos \alpha_{\vartheta}$$

where θ_{ϑ} is the coordinate of the point on dimension ϑ, γ is the distance from the origin to the point, and α_{ϑ} is the angle between the ϑ th axis and the line from the origin to the point. In m dimensions, m - 1 angles can be computed from the θ -coordinates using trigonometric

relationships. The mth angle is mathematically determined, because the sum of squared cosines must equal 1 (Reckase, 2009).

Multidimensional item difficulty (MID, or B parameter), can be interpreted like the difficulty parameter of a unidimensional model. It represents the distance and direction from the origin in the θ-space to the point of the steepest slope. B is the location that gives the ability level such that an examinee would have a .5 probability of answering the item correctly. It is the location that the item response curve discriminates the most, and consequently where the item provides the most information. This parameter has an equivalent interpretation to that of the b-parameter in UIRT models. That is, high positive values of B indicate difficult items (i.e., those that require high values of the elements of θ to yield a probability of a correct response greater than .5). Low values of B indicate items with a high probability of correct response for the levels of θ that are usually observed. This interpretation of B applies only to the direction specified by the α-vector. Thus, this analysis of the characteristics of a test item results in two descriptive measures. One is an indication of the difficulty of the test item (i.e., B) and the other is a description of the combination of the coordinate axes that is most differentiated by the test item (i.e., A). This combination is indicated by the direction of steepest slope from the origin of the θ -space. This parameter can be calculated as follows:

$$B_i = \frac{-d_i}{\sqrt{\sum_{k=1}^{m} a_{ik}^2}}.$$

Reckase (1985) proposed describing multidimensional difficulty by both the MID and the angle measure or direction cosines. Since the direction of the greatest slope from the origin with dimension k for item j=1; . . . ; J is given by

$$\alpha_k = \arccos \frac{a_{ik}}{\sqrt{\sum_{k=1}^{m} a_{im}^2}},$$

for k=1; 2; . . . ; D. The angles with the axis will reflect the patterns that are present in the discrimination parameters. Items with high discrimination/factor loadings (e.g., a_{ik}) on a factor will tend to cluster along that axis. To compare the difficulty of items, they should have similar angle measures (Yao & Schwarz, 2006). The d-parameter is not a difficulty parameter in the usual sense of a UIRT model because it does not give a unique indicator of the difficulty of the item. Instead, the negative of the intercept term divided by an element of the discrimination parameter vector gives the relative difficulty of the item related to the corresponding coordinate dimension (Reckase, 2009).

A value that is analogous to the discrimination parameter from the UIRT model can also be defined in MIRT. In UIRT, the discrimination parameter is related to the slope at the point of steepest slope for the ICC (or where the probability of a correct answer is .5). In MIRT, *A* parameter is a measure of an item's capacity to distinguish between examinees who have different locations in the factor space. If an item has a high value of *A*, then it will provide a relatively large amount of information somewhere in the factor/trait space. Because the a-parameters are related to the slope of the surface and the rate of change of the probability with respect to the coordinate axes, the *A*-parameter is usually called the slope or discrimination parameter. The equivalent conceptualization for the discrimination parameter in the MIRT case

is the slope of the item response surface at the point of steepest slope in the direction from the origin of the θ -space. This results in the multidimensional discrimination index of

$$A_i = \sqrt{\sum_{k=1}^{m} a_{ik}^2}$$

where m is the dimension and a is a vector of dimension m of item discrimination parameters (Reckase, 2009).

For the dichotomously scored item in the Rasch multidimensional model, the probability of a response of an individual j on an item i is as follows:

$$P\left(U_{ij} = 1 \,|a_i, d_i, \theta_j\right) = \frac{e^{a_i\,\theta_j\prime + d_i}}{1 + e^{a_i\,\theta_j\prime + d_i}},$$

where a_i is a vector such that $a_i = b_{ik}$, and d_i is a scalar value equal to $a'_{ik}\varepsilon.$ a_i is a characteristic of Item i that is in this model specified by the test developer. The general form of the exponent can be given as follows, when all the *a* parameters are set equal to the same value (e.g. a_{i*} :

$$a_i\,\theta'_j + d_i = a_i\,\theta'_j + d_i = a_{i1}\,\theta_{j1} + a_{i2}\,\theta_{j2} + \cdots + a_{im}\,\theta_{jm} + d_i = a_{i*}\left(\theta_{j1} + \theta_{j2} + \cdots + \theta_{jm}\right) + d_i.$$

For the 2PL MIRT extension,

$$P\left(U_{ij} = 1 \,|a_i, d_i, \theta_j\right) = \frac{e^{a_i\,\theta_j\prime + d_i}}{1 + e^{a_i\,\theta_j\prime + d_i}}.$$

The exponent of *e* in this model can be expanded to show the way that the elements of the *a* and θ vectors interact.

$$a_i\,\theta_j' + d_i = a_{i1}\,\theta_{j1} + a_{i2}\,\theta_{j2} + \cdots + a_{im}\,\theta_{jm} + d_i = \sum_{l=1}^{m} a_{il}\,\theta_{jl} + d_i .$$

The exponent is a linear function of the elements of θ with the *d* parameter as the intercept term and the elements of the *a*-vector as slope parameters. The expression in the exponent defines a line in an m-dimensional space. While the formulas for M1PL and M2PL appear to be the same, the difference is that while in the Rasch MIRT the *a* parameter is preset by the test developer, in M2PL it is estimated from the data.

A M3PL model is a fairly straightforward extension of the M2PL model provides for the possibility of a non-zero lower asymptote to the model. This is a multidimensional extension of the three parameter logistic UIRT model:

$$P\left(U_{ij} = 1 \,\middle|\, a_i, c_i, d_i, \theta_j\right) = c_i + (1 - c_i)\frac{e^{a_i\,\theta_{j'} + d_i}}{1 + e^{a_i\,\theta_{j'} + d_i}}$$

with the exponent similar to the M2PL model (Reckase, 2009).

The concept of item information that is used in UIRT can also be generalized to the multidimensional case. However, at each point in the θ-space, the slope of the multidimensional item response surface differs depending on the direction of movement from the point. The item information can be expressed by the following formula:

$$I_\alpha(\theta) = \frac{[\nabla_\alpha P(\theta)]^2}{P(\theta)Q(\theta)},$$

where α is the vector of angles with the coordinate axes that defines the direction taken from the θ -point, ∇_α is the directional derivative or gradient, in the direction α (figure 10). For a M2PL model, ∇_α is calculated as follows:

$$\nabla_\alpha P(\theta) = a_1 P(\theta)Q(\theta)\cos\alpha_1 + a_2 P(\theta)Q(\theta)\cos\alpha_2 + \cdots +$$

$$a_m P(\theta)Q(\theta)\cos\alpha_m = P(\theta)Q(\theta)\sum_{\vartheta=1}^{m} a_\vartheta \cos\alpha_\vartheta$$

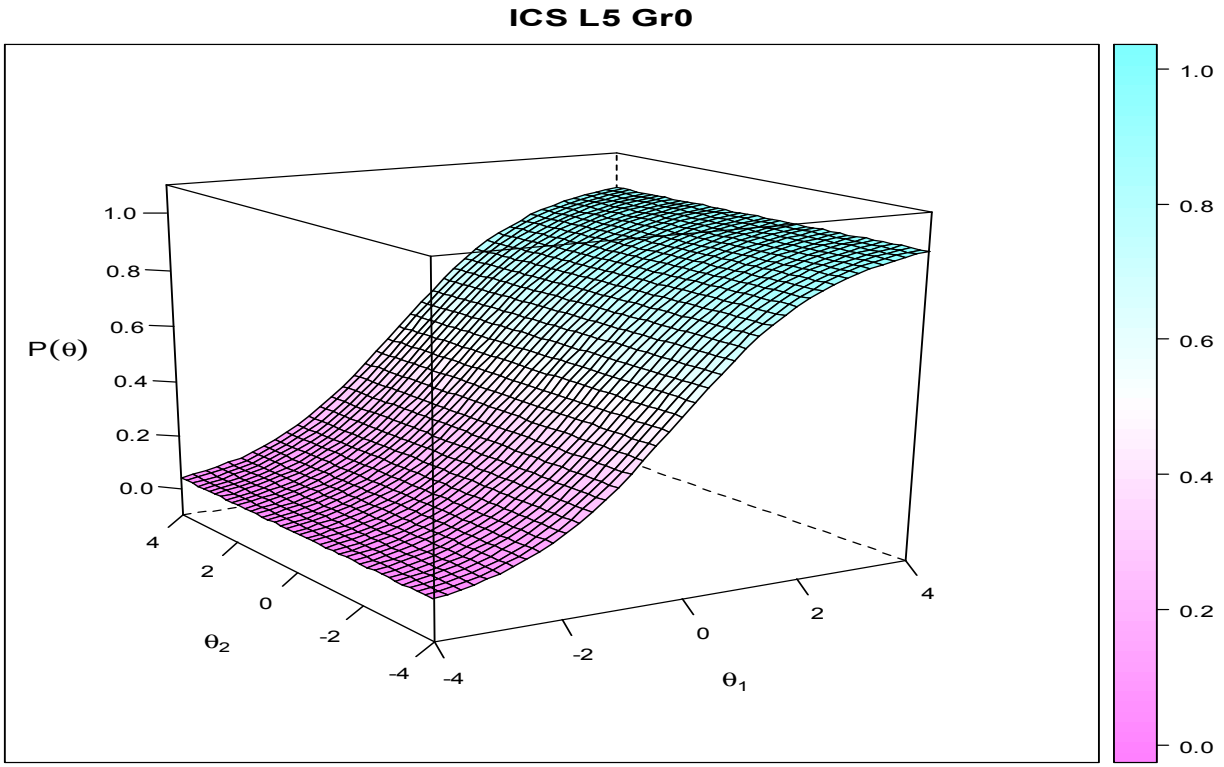

Figure 10. MIRT item information surface for a 2PL model.

The equivalent of test characteristic curve in UIRT is the test characteristic surface (TCS) in MIRT. The expression for the TCS is very similar to that of UIRT, except that the expectation is conditioned on the θ-vector instead of the unidimensional value of θ. It can be computed as follows:

$$E(y_j|\theta_j) = E\left(\sum_{i=1}^{n} u_{ij}|\,\theta_j\right) = \sum_{i=1}^{n} E\left(u_{ij}\,|\theta_j\right) = \sum_{i=1}^{n} P\left(u_{ij}|\theta_j\right).$$

The TCS is simply the sum of the item characteristic surfaces for the items in the test (Reckase, 2009).

MIRT models can be viewed as exploratory or confirmatory. Exploratory MIRT models specify two or more dimensions that are combined to predict item responses. Exploratory MIRT models are sometimes described as full information factor analysis models. The goal of this type of model is similar to factor analysis: to determine the number and nature of the factors that underlie item performance. Like in factor analysis, the nature of the dimensions is interpreted by the relative pattern of discriminations across the items. The information, however, is extracted not only from the correlations between items, like in factor analysis, but from the response level data (Embretson & Reise, 2000). The exploratory analysis of item response data is used when either there is no clear hypothesis for the structure of the item response data, or when an unconstrained solution is seen as a strong test of the hypothesized structure (Reckase, 2009).

Confirmatory analyses require a clear hypothesis for the structure of the item response data. That is, there must be hypotheses about the number of coordinate dimensions needed to model the data and the relationship of the item characteristic surface to the coordinate the axes. The types of confirmatory analyses that are typically done specify the relationship of the direction best measured by a test item (the direction of maximum discrimination) with the coordinate axes. In many cases, what is labeled as an exploratory analysis also has a confirmatory analysis component, because the number of coordinate dimensions is selected prior to estimating the item and person-parameters (Reckase, 2009). Various confirmatory models could be investigated by associating items with specific dimensions or traits (Yao & Schwarz,

2006). For example, one assumption may be that multiple-choice and constructed-response items measure different traits or dimensions. On the other hand, a hypothesis about the number of dimensions can be based on the number of subscales or standards that the test intends to measure according to the blueprint.

Both exploratory and confirmatory models may be compensatory or non-compensatory. According to Reckase (2009), these model types are defined by the way the information from a vector of θ-coordinates is combined with item characteristics to specify the probability of responses to the item. Compensatory models are based on a linear combination of θ -coordinates. The linear combination of θ –coordinates can yield the same sum with various combinations of θ -values. If one θ -coordinate is low, the sum will be the same if another θ -coordinate is sufficiently high. The compensatory models are more consistent with a more holistic view of the interaction of persons and test items, because usually persons bring all of their skills and knowledge to bear on all aspects of the items.

In non-compensatory models, the cognitive tasks in a test item are separated into different parts, and a unidimensional model is used for each part. In other words, one has to have mastery of all aspects that go into being able to answer an item to be able to respond correctly. The probability of correct response for the item is therefore the product of the probabilities for each part. The use of the product of probabilities results in nonlinear features for this class of models. Also, the fact that the probability of correct response cannot exceed the highest of the probabilities in the product reduces the compensation of a high θ -coordinate for a low θ – coordinate (Reckase, 2009).

Similarly to UIRT, MIRT models can be scored in more than one category. The development of polytomous MIRT models is relatively new (Reckase, 2009). These polytomous

models are normally compensatory. One example of such polytomous models is the multidimensional extension of the generalized partial credit (MGPC) model, which is designed to describe the interaction of persons with items that are scored with more than two categories. The maximum score for Item i is represented by K_i. To be consistent with the way dichotomous items are scored, the lowest score is assumed to be 0 and there are $K_i + 1$ score categories overall. The score assigned to a person on the item is represented by k = 0,1, …, K_i. Mathematically the MPCG model can be expressed as follows:

$$P\left(u_j = k \middle| \theta_j\right) = \frac{e^{ka_i\theta'_j - \sum_{u=1}^{k}\beta_{iu}}}{\sum_{\vartheta=0}^{K_i} e^{\vartheta a_i\theta'_j - \sum_{u=0}^{\vartheta}\beta_{iu}}},$$

where k is a score of person j in the category u, given the person's θ; a_i is the item discrimination parameter; β_{iu} is the threshold parameter of item i for score category *u*, β_{i0} is defined to be 0, ϑ is the number of steps in an item; and K_i is the highest score that can be obtained on item i.

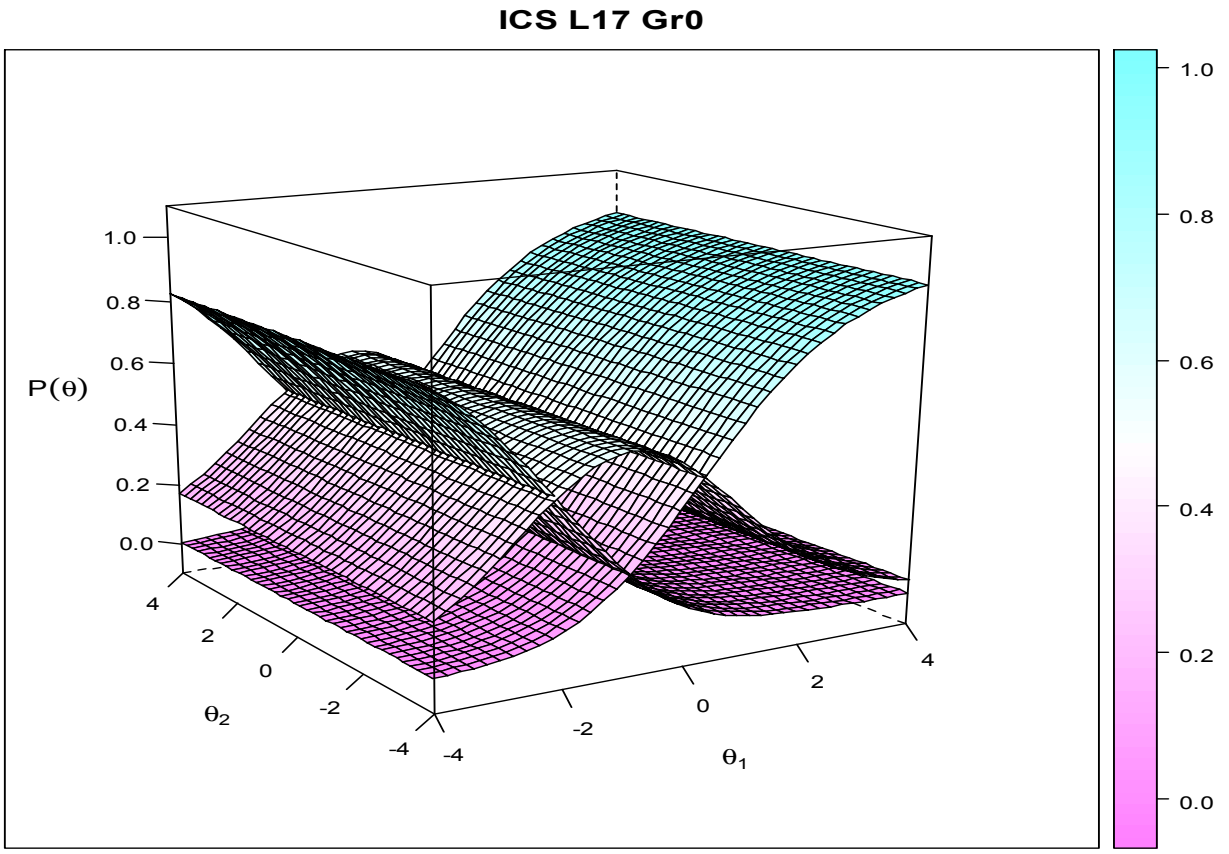

Figure 11. Item characteristic surface for a polytomous item scored in 3 categories with a 2 PL multidimensional GPCM model.

There are two important differences between the equation for the MGPC model and that for the unidimensional GPC model. First, the model does not include separate difficulty and threshold parameters. Second, because θ is a vector and the β's are scalars, it is not possible to subtract the threshold parameter from θ. Instead, the slope/intercept form of the GPCM is used as the basis of the multidimensional generalization, $a\,\theta + d$, but with the sign of the intercept reversed. The result is that the β's cannot be interpreted in the same way as the threshold parameters in the UIRT version of the model. To determine the point where the probabilities of obtaining the adjacent scores are equal, one needs to solve the equation

$$ka_i\theta_j' - \sum_{u=1}^{k}\beta_{iu} = (k+1)a_i\theta_j' - \sum_{u=0}^{k+1}\beta_{iu}.$$

This is the equation for a line in the m-dimensional space used to represent the item. The only part of this expression that changes for different adjacent score categories is the intercept term, β. As was the case for the UIRT version of the generalized partial credit model, this parameter controls the location of the thresholds between score categories. The MGPC is a compensatory model in the same sense as the UIRT version of the model in that a high value on θ_ϑ can compensate for a low value on θ_ω resulting in a high expected score on the item.

Similarly to score estimates in the UIRT, there are Bayesian and non-Bayesian methods for the subscore estimates by MIRT models. MAP and EAP perform similarly, however, EAP takes more time than MAP; for EAP, as the number of quadrature points increases, the time increases nonlinearly. With the development of MCMC technique, the abilities can be derived by MCMC sampling for the posterior distribution. The length of the test, the number of items in each subscale, and the population distributions are clearly factors that affect the accuracy of the estimation (Yao, 2010). Several software programs are capable of MIRT score estimation, including flexMIRT (Houts & Cai, 2013) and BMIRT (Yao, 2003).

Reliability Estimation within MIRT Framework

Similarly to UIRT, reliability in MIRT depends on item information. As was the case for the unidimensional IRT models, the reliability of a test can be determined by summing the information available from each item. However, in MIRT models the sum of the information estimates must be for the same direction in the θ -space. The test information function at point $\vec{\theta}$ in θ space in the direction $\vec{\alpha}$ can be obtained by

$$I(\theta_\alpha) = \frac{1}{Var(\hat{\theta}_\alpha)}.$$

Consequently, similar to the UIRT reliability estimation,

$Var(\hat{\theta}_\alpha) = \frac{1}{I(\theta_\alpha)}$, and

SEM $(\theta_\alpha) = \sqrt{Var(\hat{\theta}_\alpha)}$.

The SEM for each domain can be derived by using the angle for that dimension to be 0, and all other angles 90° (Yao, 2010). He (2009) outlines similar steps for measuring error for both unidimensional and multidimensional IRT. Specifically, the errors of the composites of ability measures in different domains are calculated separately, and then if desired, the composite reliability may be calculated. CSEM in MIRT is calculated similarly to UIRT, since SEM is calculated at specific levels of θ .

Subscore and Reliability Estimation of Augmented Methods

The idea behind using an augmentation procedure is to borrow information from some other source collateral to examinee responses in order to reduce error. Technically, the information can be borrowed from various sources, such as in-test sources (e.g. the student's total score or means of group scores) and out-of-test sources (e.g. student demographic information) (Mislevy, 1987).

The operating principle behind such an approach is that the precision of an estimate can be improved by regressing that estimate toward some aggregate value, typically the mean of collection of values. Through such procedures, the reliability of each subscore is augmented by obtaining information from the other portions of the test and using these data to stabilize the

estimate of the ability suggested by the subscore (Wainer, et al., 2001). There are three basic approaches that have been suggested for how to implement this information: (1) using Bayesian IRT ability estimates for these subscores that incorporate an informative prior to stabilize the subscore ability estimate, (2) using regressed estimates of ability, or linear combinations of scores, to create more reliable subscore estimates, and (3) using non-linear combinations of ability estimates, as in the case of a multidimensional IRT (MIRT) modeling approach (Skorupski, 2008). While the subscore reliability normally increases after an augmentation, one issue with producing augmented scores is that they may not provide much distinct information in subscores that is not reflected in total scores. When augmented scores are obtained in such a context, they are essentially replications of the total score and serve little diagnostic purpose (Sinharay et al., 2007).

Wainer et al.'s (2001) Augmentation Method: CTT Application

Based on Kelley's regression method that weighs the observed subscores based on the group mean ($\hat{\tau} = \rho\chi + (1-\rho)\,\mu$, where $\hat{\tau}$ represents an estimate of true score (τ), χ is the observed score, μ is the group mean, and ρ is the reliability of the test), Wainer et al. (2001) suggested a procedure that is essentially a multivariate estimation of subscores, which can be expressed in the following way:

$\hat{\tau} = \bar{X} + B\,(x - \bar{X})$,

where **B** is a matrix of weights, which is analogous to the estimated reliability for a single measure, $\bar{X}$ is the group mean on a given subscale, and x is the observed score of an individual on a given subscale. The matrix **B** contains weights that combines the observed scores in x into estimates of the true scores in τ. If **B=I**, that is, **B** is an identity matrix with 1's on the diagonal

and 0's in off-diagonal positions, the scores are perfectly reliable and the observed scores are the same as the estimated scores. If **B=0**, that is, all values in the matrix equal 0, all observed scores are regressed to the mean $\bar{X}$. While Kelley's regression only uses the group mean and reliability of the score to regress subscore values to, Wainer's method also uses correlations between subscores to augment the subscores.

The **B** matrix may be determined directly by multiplying the matrix of covariances of observed scores $\mathbf{S^{obs}}$ by the inverse of the matrix of covariances of true scores $\mathbf{S^{true}}$**:**

$\mathbf{B = S^{obs} (S^{true})^{-1}}$**.**

The diagonal elements of $\mathbf{S^{obs}}$ contain the observed variances of subscores, and off-diagonal elements – the covariances between observed subscores. These off-diagonal elements are used as the off-diagonal elements of the estimates of $\mathbf{S^{true}}$. The covariance matrix of true subscores is then estimated as follows:

$S_{xx\prime}^{true} = S_{xx\prime}^{obs}$ for x ≠ x';

$S_{xx\prime}^{true} = \rho_{xx\prime}\ S_{xx}^{obs}$ for x=x'.

The **B** values are therefore the weights for the linear combination of the deviation scores that predict the best estimates for the subscale scores. As mentioned previously, in the absence of any error, the true and raw covariance matrices coincide, and the observed score equals the estimated score. As error increases, the term $B\ (x - \bar{X})$ shrinks toward 0, so the best estimate of true score becomes the group mean. Therefore the B coefficient depends on the covariances between subscores, as well as the reliability of subscales.

If for a given test we have the information about variances of observed subscores $x_1, x_2, \ldots, x_n$, covariances between pairs of observed subscores, the group means of subscores, and the reliability of subscales, we will be able to estimate the parameters (β's) in the equation

$$\hat{\tau} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n$$

By solving the matrix equation $\mathbf{B} = \mathbf{S}^{obs}(\mathbf{S}^{true})^{-1}$ and obtaining the β weights for all the subscales. We will get the **B** matrix of weights that are used to combine the observed subscale scores into the regressed estimates of true scores. Thus, $\bar{\tau}$ becomes an improved estimate of the subscale true score, based on an examinee's performance for that subscale, as well as his or her performance on all other subscales. In sum, Wainer, et al.'s (2001) proposed method posits that an improved estimate for a subscale true score, $\hat{\tau}$, may be obtained for each examinee by combining information from the observed subscore with a simple linear combination of other observed subscores (Skorupski, 2008).

Wainer et al.'s (2001) Augmentation Method: IRT Application

Wainer et al.'s (2001) subscore augmentation can use either the number-correct scores or the IRT scale score estimates as the observed score. When the IRT approach is used, subscore augmentation may be thought of as a multi-stage estimation procedure for proficiency estimates for the domains (Thissen & Edwards, 2005). In the first stage, unidimensional IRT ability estimates are obtained using, for example, one of maximum likelihood (MLE), maximum a posteriori (MAP), or expected a posteriori (EAP) methods. The values of MLE(θ), EAP(θ), and MAP(θ) in the augmentation procedure correspond to the regressed estimates in Kelley's method, with one difference: Kelley's regressed estimates of the true scores shrink linearly, or proportionally, towards the mean, whereas the amount that EAP(θ) and MAP(θ) shrink towards the mean is proportional to the score variance. That is, the values of EAP(θ) and MAP(θ) shrink more toward the mean when they are associated with response patterns that provide relatively

less information, and they shrink less when the response patterns provide more information. For that reason, to be able to treat the estimated IRT scores the same way we treated the observed scores in the regression formula, we will need to treat the error of measurement for the IRT scale scores as average across all levels of θ, similar to the CTT approach. In addition, we will need to make a correction to the IRT scale scores to remove their regression towards the mean.

After the values of EAP(θ) or MAP(θ) have been computed, a formula analogous to Kelley's regression formula can be applied:

$MAP\ [\theta_v] \approx\ \rho_v\ MAP * [\theta_v]$,

Where ρ_v is an estimate of the reliability of subscale v, and $MAP * [\theta_v]$ is a hypothetical IRT scale estimate of θ_v that is not regressed toward the mean; that is, the values of $MAP * [\theta_v]$ are like the observed summed scores (Wainer et al., 2001). In addition, it is assumed that the group mean is 0. Therefore to calculate the value of $MAP * [\theta_v]$,

$MAP * [\theta_v] = \frac{MAP\ [\theta_v]}{\rho_v}$.

To compute $MAP * [\theta_v]$ for each subscale, we need some estimate of reliability of the IRT scale score for that subscale, ρ_v. Wainer et al. (2001) suggest using the following formula to do this:

$\rho_v = \frac{Variance\ [MAP[\theta_v]]}{Variance\ [MAP[\theta_v]] + Average\ [SE^2[\theta_v]]}$,

In which $Variance\ [MAP[\theta_v]]$ is the variance of the IRT scale scores for subscale v, and $Average\ [SE^2[\theta_v]]$ is the average value of the variance of the error of measurement associated with those scores. This formula is used by Wainer et al. (2001) to calculate non-augmented IRT reliability scores; at the same time, the formula is necessary to calculate the scores that are

conceptually equivalent to observed IRT scores, which serve as an intermediary step in the calculation of augmented IRT scores.

After computing an estimate of ρ_v for each subscale using the equation above, we can compute $MAP * [\theta_v]$ for each examinee for each subscale, and then compute $\mathbf{S}^{\mathbf{obs}}$ as the covariance matrix between the scale scores $MAP * [\theta_v]$. Once we have estimated $\mathbf{S}^{\mathbf{obs}}$ and ρ_v, we can compute $\mathbf{S}^{\mathbf{true}}$ by subtracting from $\mathbf{S}^{\mathbf{obs}}$ the diagonal matrix **D**, the diagonal elements of which are estimates of error variance. Error variance for each subscale is estimated as follows:

$$\sigma_{ev}^2 = \sigma_{xobs}^2 * (1 - \rho_v).$$

Now we can use the equation

$$\mathbf{S}^{\mathbf{true}} = \mathbf{S}^{\mathbf{obs}} - \mathbf{D}$$

to compute $\mathbf{S}^{\mathbf{true}}$, and then compute the empirical Bayes estimates of the IRT scale scores as follows:

$$\widehat{\underline{MAP}}[\theta] = \underline{MAP}[\theta]. + S^{true}\,(S^{obs})^{-1}(\underline{MAP}*[\theta]_j - \underline{MAP}[\theta].) =$$

$$\underline{MAP}[\theta]. + B * (\underline{MAP} * [\theta]_j - \underline{MAP}[\theta].)$$

where $\underline{MAP} * [\theta]_j$ is the vector of IRT scale scores for the subscales for examinee j; $\underline{MAP}[\theta]$. Is the average vector of IRT subscale scores; $\widehat{\underline{MAP}}[\theta]$ is the vector of empirical Bayes estimates of the IRT scale scores, and

$$\boldsymbol{B} * = \boldsymbol{S}^{\boldsymbol{true}}\,(\boldsymbol{S}^{\boldsymbol{obs}})^{-\mathbf{1}}.$$

The weights in the matrix **B*** are the coefficients for the values of $MAP * [\theta_v]$ for each examinee. Because $MAP * [\theta_v]$ is computed from $MAP[\theta_v]$ by dividing each $MAP[\theta_v]$ by the

corresponding subscale reliability estimate, we can similarly obtain weights B by dividing **B*** by the reliability of the corresponding subscale:

$B = \frac{B*_{vv\prime}}{\rho_{v\prime}}$.

Then the weights B are used to compute the augmented estimates from the original values of $\underline{MAP}[\theta]$:

$\widehat{\underline{MAP}}[\theta] = \underline{MAP}[\theta]. + B(\underline{MAP}[\theta]_j - \underline{MAP}[\theta].)$

Reliability Estimation for Augmented IRT Subscores

Wainer et al. (2001) propose that an estimate of augmented subscale reliability can be derived by calculating the ratio of the unconditional true score variance to unconditional estimated true score variance. The numerator of this fraction is the unconditional true score variance of the vth subscale, which is the *v*th diagonal element of the matrix

$A = S^{true}\,(S^{obs})^{-1}S^{true}\,(S^{obs})^{-1}S^{true}$.

The denominator, which is the unconditional score variance of the estimates for the *v*th scale, is the *v*th element of the matrix

$C = S^{true}\,(S^{obs})^{-1}S^{true}$.

Therefore the reliability of an augmented subscale can be expressed as follows:

$r_v^2 = a_{vv}/\,c_{vv}$.

The value of SEM can then be calculated with the formula analogous to the SEM for CTT-derived scores:

SEM = $\sqrt{Variance\,[MAP[\theta_v]]} * \sqrt{1 - r_v}$.

Since Lord's formula for CSEM calculation for CTT scores is based on the number of points possible, there are no suggestions that it can be applied in the IRT environment. Only the average SEM is available for the augmented IRT scores.

Studies Comparing Reliability of Different Subscore Reporting Methods

A number of studies compared the approaches to subscore reporting and methods to increase subscore reliability mentioned above. Studies comparing augmented vs. non-augmented methods of reporting, CTT vs, IRT, and unidimensional IRT vs. MIRT have been done. These studies compared different combinations of subscore calculation methods, and used different statistical methods for these comparisons; however, their findings are consistent in comparing these different methods.

CTT vs. IRT Methods

A number of studies suggest that item response theory trait estimates should be used in analyses rather than number right (NR) or summated scale (SS) scores due to the fact that IRT subscores have higher reliability and precision (Yao & Boughton, 2007; Luecht, 2003; Shin, 2007). For example, Shin (2007) found that proportion correct score performs worse than UIRT, OPI, and Wainer et al.'s (2001) regression method. Wainer's method performed better than the OPI. Dwyer et al. (2006) found that percent correct subscores have larger root mean square errors (RMSE) than OPI, Wainer et al.'s method, and MIRT. Haberman and Sinharay (2010) found that MIRT-derived subscores and subscores augmented using Haberman's (2008) method were significantly more reliable than raw scores. Thissen and Orlando (2001) stated that IRT scaling tends to produce trait estimates that are linearly related to the underlying trait being

measured. Therefore, IRT trait estimates can be more useful than summated scores when examining relationships between test scores and external variables. Ferrando and Chico (2007) stressed that the increased test information afforded by IRT models versus CTT applications implies the possibility of greater accuracy or precision in trait estimates versus CTT-based scores. IRT models offer more flexibility in scaling sets of items with the same response scale. While CTT makes an assumption that all items are equally related to the trait being measured, IRT can address some of these constraints (Xu & Stone, 2012). Of course, these advantages of IRT models are only realized if the IRT model fits the data.

MIRT vs. UIRT Methods

Many studies have found that MIRT has an advantage in reliability and precision over other methods of subscore reporting. For example, Wang, Chen and Cheng (2004) investigated how measurement precision can be improved through use of the correlations between latent traits. They compared subscore estimation with a unidimensional and multidimensional approach using EAP estimation. They found that the multidimensional approach improves measurement precision dramatically; the higher the correlations between subscores, the greater the number of latent traits, and the shorter the subtests, the more significant the improvement in reliability will be as compared with UIRT. When tests are long, the improvement shows a ceiling effect. In other words, there is less to be gained using collateral information when the target latent trait is already well determined by its test score alone. The direct estimation of the variance– covariance matrix (and the correlation matrix) of latent traits is another advantage of the multidimensional approach. When UIRT is used, the correlations between latent traits cannot be estimated directly. Even though a two-step procedure can be used to estimate the attenuated correlations, and then to adjust them for the attenuation caused by imperfect reliability, such estimation of the correlation

matrix has been found to be biased. In contrast, direct estimation of the variance–covariance matrix of latent traits via the multidimensional approach obviates the problem of the bias introduced by the two-step procedure.

Yao and Boughton (2007) compared the reliability of percentage correct on subscale number-correct scores (NC), a unidimensional Bayesian IRT scoring method, and several Bayesian MIRT scoring methods. They found that the unidimensional scoring method and multidimensional IRT Bayesian subscale score approach performed similarly when the correlation was low or close to .0, but as the correlation increased, the Bayes MIRT method performed better, whereas the unidimensional method's performance remained the same. Similarly, Yao and Boughton (2007) compared a MIRT approach using MCMC estimation to the objective performance index (OPI; Yen, 1987) and found that as the correlation increased across the subscales, the OPI procedure produced more biased results, while the MIRT approach produced more reliable and robust subscores across all conditions studied.

Boughton, Yao, and Lewis (2006) studied the performance of an MIRT model under several conditions: varying sample size, correlation between subscales, the number of items contributing to each subscale, and whether items contributed only to a single subscale (simple structure) or to several subscales (complex structure). In general, as the correlation increased, the error in parameter estimation also increased for all of the tests with complex structured items, but decreased for tests composed only of simple structure items, with the largest increase being for the 0.8 correlation condition. It was found that at least 10 to 12 items are needed to produce stable item parameters estimates for a given subscale, regardless of population correlation for subscales with simple structure. Ability estimation improved as the correlation increased to 0.8. The sample size, however, did not improve ability estimation.

De la Torre and Song (2009) compared UIRT and a higher-order IRT (HO-IRT) approach (in which there is a higher order factor of the overall ability, but it has no direct effect on the examinee performance; the performance variability is accounted for solely by the specific domain factors or abilities) to subscore estimation. The domain ability estimates using the CU-IRT method were identical to the HO-IRT estimates when the correlation between domains is 0. UIRT analysis of tests that estimates the overall ability using all the items in a test still ignores its multidimensionality; under this framework, domain abilities are estimated by repeating the UIRT approach multiple times using subsets of the items. Therefore, depending on the disparity of the domains, the general or overall ability estimate may not be valid depending on the extent to which the unidimensional assumption is violated. And, with the correlation between the domain abilities ignored, the multiple applications of the UIRT approach are suboptimal, and provide domain ability estimates that are unreliable, especially when the number of items in each domain is small. They found smaller MSE and higher reliability when subscales were longer, or when more domains measuring abilities with at least moderately high correlations were considered; the improvements due to test length tapered off when longer tests were involved. Although HO-IRT estimation is expected to provide better estimates than UIRT estimation, the degree of improvement can be negligible when the correlations between the domain abilities are low or when the domain abilities have already been well estimated using long tests. Finally, based on the correlation values, the sample size had no noticeable impact on the quality of the ability estimates.

De la Torre, Song, and Hong (2011) compared four subscoring methods - multidimensional scoring (MS), augmented scoring (AS), higher order item response model scoring (HO), and objective performance index scoring (OPI)—by examining how test length, number of subtests or domains, and correlation between the abilities affect the subtest ability

estimation. The correlation-based methods (i.e., MS, AS, and HO) provided largely similar results, and performed best under conditions involving multiple short subtests and highly correlated abilities. In most of the conditions considered, the OPI method performed poorer compared to other methods on both ability estimates and proportion correct scores.

HO and MS scoring was based on MCMC estimation, while AS and OPI scoring was based on EAP estimation. As would be expected, in methods using EAP, shrinkage in estimates was observed, thus resulting in EAP estimates with smaller SD than the true ability. The relative size of the shrinkage was more apparent for shorter tests, but, as expected, not related to the number of and correlations between dimensions. Methods that incorporate the correlational information among the tests (i.e., HO, MS, and AS) produced estimates with smaller shrinkage, and their impact was more apparent for shorter but highly correlated tests. The SD of the OPI ability estimate was larger, and in some cases, much larger than the SD of true θ.

As expected, better estimates can be obtained with longer tests, but the differential benefit of adding a fixed number of items diminishes with higher test reliability. In general, the improvement over the EAP estimates using the three subscoring methods was greater when correlation between domains was higher, but the number of items was lower. In many cases, however, the performance of HO, MS, and AS methods was pretty comparable. This study also shows that although EAP estimates may have smaller bias and standard error, further improvement can be achieved by taking into account the correlational structure of the abilities.

This is one of the few studies that addresses the reliability of different subscoring methods for examinees at different ability levels. While the augmentation methods' reliability was pretty similar, some discrepancies between these methods can be observed in the measurement of extreme abilities—HO had the smallest mean absolute deviation of ability estimate (MAD), followed by MS, then by AS. Compared to the correlational methods, OPI

yielded the largest MAD for all but the extremely low-ability examinees; more importantly, OPI was worse than other methods for abilities in the middle range of the continuum where most of the examinees are typically located. The medians (i.e., 50th percentile) of the four subscoring methods were not very different. For both measures of central tendency, HO and MS showed a more similar pattern of over- and underestimation compared to the AS and OPI estimates. At the 95th percentile, the four subscoring methods showed identical or almost identical estimates of ability. The correlation-based methods (i.e., HO, MS, and AS) gave highly comparable results across the different conditions except for extreme abilities where HO and MS may perform better. Overall, the largest discrepancies between methods were found at the lowest spectrum of abilities (5^{th} percentile).

A study by De la Torre and Patz (2005) compares the EAP estimates of augmented UIRT and MIRT approaches. The factors investigated in this article are: (a) the number of abilities, (b) the number of items, and (c) the degree of correlation between the abilities. They found that when only two abilities are concurrently considered, the efficiency of the EAP-M method was not evident unless the abilities are very highly correlated (i.e., p = .90). When more dimensions were simultaneously used, the efficiency of the EAP-M method was evident for abilities that were reasonably highly correlated (i .e., p = .70). Efficiency ranged from 1.16 to 1.95. For 10-item tests where five abilities with p = .90 were simultaneously estimated, the precision of the EAP-M estimates was equivalent to the precision of the EAP-U estimates obtained from tests twice as long. For other conditions, depending on the original test length the additional precision was equivalent to adding 4 to 26 items to the test. The increase in precision from increasing the number of abilities was less evident when long tests were used.

The MIRT method gives the same results as the unidimensional approach when abilities are uncorrelated. However, when abilities are correlated, taking the correlation into account can

lead to noticeable improvements in ability estimates, especially when there are multiple short tests and the underlying correlation is high. Among several methods of ability estimation, EAP-U has been preferred for its small bias and standard error (Kim & Nicewander, 1993; Thissen & Orlando, 2001). But, as the results of this article have shown, employing simultaneous estimation can further reduce the bias and standard error of the estimates.

De la Torre and Patz (2005) find that it would be advantageous to use MIRT: although they concur that some of the improvements from using this approach are relatively modest, it can be achieved without much additional cost. The score can be made more reliable, or, given a desired level of reliability, the number of items can be reduced without loss of accuracy. These authors stress the importance of the decision as to which method of reporting to use by the purpose of subscores. For example, if more accurate score profiles lead to more efficient diagnosis or more precise targeting of instructional resources, then their use could be supported. Finally, we observe that multidimensional scores may serve to complement rather than to replace traditional scoring of test batteries. Traditional unidimensional scores may be reported at the domain level (e.g., scale scores and their associated norm-referenced and/or criterion-referenced derived scores), and multidimensional scores could be used to inform finer-grained reporting, such as skills profiles and objective-level scores. Traditional approaches to this type of fine-grain reporting suffer from insufficient reliability because of the small numbers of items associated with each fine-grain reporting category.

Dwyer et al. (2006) compared OPI, Wainer et al.'s (2001) regressed true subscores; MIRT, and percent correct subscores. They found that the MIRT and Wainer et al.'s subscores were more accurate than the OPI and percent correct subscores, and the OPI produced better

subscore estimates than did percent correct. Overall, MIRT provided the best predictions of ability at the subscale level, followed closely by Augment and OPI.

DeMars (2005) compared two MIRT models (a bifactor model, and a two-factor model where the traits measured by the subtests were separate but correlated) two UIRT models (one in which each subtest borrowed information from the other subtest, and the other where subscores were treated as being independent). The independent unidimensional scores showed the greatest bias and RMSE; the relative performance of the other three methods varied with the subscale. The bifactor and two-factor MIRT models showed very similar levels of bias and RMSE: on one test higher than the augmented UIRT scores at the extremes, and on the other test lower.

Haberman and Sinharay (2010) compare MIRT, augmented subscores using Haberman's (2008) augmentation, and CTT-derived subscores from the standpoint of whether they have added value over the reporting of total score. They found that the MIRT subscores almost always yield a proportional reduction in mean squared error (PRMSE) at least as high as those provided by the augmented subscores. The differences are often quite small, but they are appreciable in a number of cases. They also found, similarly to De la Torre and Patz (2005), that MIRT subscores are very close to regression-based augmented subscores.

Other studies, however, found that unidimensional and multidimensional methods can often have very comparable results. For example, Gessaroli (2004) used multidimensional IRT (MIRT) to compute the objective score. He then compared it to the Wainer's EB method (Wainer et al., 2001), and found that the EB method had almost the same results as the MIRT methods. Additionally, some researchers suggest that MIRT methods, while producing very similar results, are more computationally demanding and are harder to interpret. For example, Luecht (2003) compared number correct subscores, Bayes EAP estimate based on the UIRT total-test (UIRT-T) item calibration; Bayes MAP estimate based on the UIRT separate calibrations of items for the

separate competency areas (UIRT-S); and MAP scores based on a multidimensional calibration of the entire test, with one factor representing each competency area (MIRT) using the full-information, common factor solution with a varimax rotation. The UIRT-T method was selected because it offers the advantage of requiring only a single calibration of the items. On the downside, the UIRT(T) method loses some of the unique variance associated with the individual factors and can induce an upward correlation bias between the subscores. In contrast, the UIRT(S) approach treats each competency area as a separate subtest. At the same time, the independent calibrations allow the trait composites for that the individual competency metrics to diverge from each other, as necessary. It was found that UIRT(S) scores appear to provide the most accurate estimates. The MIRT results are likewise reasonable, but hardly seem to justify the added complexity of using a multidimensional model. In both cases, the two unidimensional methods, UIRT(T) and UIRT(S), produced diagnostic subscore profiles that were more consistent with the true profiles.

A number of researchers characterize MIRT as producing an improvement in the reliability of subscore estimation that comes at the cost of more complex interpretation and higher computational demands (Stone et al., 2010, Luecht, 2003; Gessaroli, 2004; Haberman & Sinharay, 2010; Yao & Boughton, 2007). Haberman and Sinharay (2010) also mention that testing programs with limited time for data analysis may find MIRT less attractive of an option. In addition, use of the MIRT-based approach results in estimates that are more difficult to explain than are raw scores, although this issue can be alleviated by alternative scalings.

MIRT methods are seldom used in large scale assessments (Shin, 2007). However, an argument for using MIRT is that if necessary, both domain and total scores can be computed all at once; in addition, this approach is more true to the structure of the construct, that is more

likely than not multidimensional. Therefore it approaches the construct more realistically and increases the strength of the validity argument for the test that comes from the content structure.

Augmented vs. Non-augmented Methods

A number of studies (e.g., Wainer et al., 2001; Puhan et al., 2010; Sinharay, 2010; Skorupski & Carvajal, 2010; Skorupski, 2008) have shown that augmented subscores and weighted averages often are substantially more reliable than non-augmented subscores, and often have added value over simple subscores and the total score. For example, Dwyer et al 2006 found that Wainer et al.'s (2001) regression-based augmentation procedure performed only slightly worse from the standpoint of reliability than MIRT, and better than OPI and percent correct subscores. Edwards and Vevea (2006) examine the use of augmentation in two popular scoring methods: summed scores and IRT scale scores for summed scores. They found that the subscores produced by the EB augmentation procedure represent a global improvement over nonaugmented subscores. For both the IRT-based EAPs and for summed scores, the augmented scores exhibit smaller RMSE, correctly place a higher percentage of simulees in appropriate ability groups, and are more reliable. The magnitude of the improvement gained through the use of the augmentation procedure is a function of correlation between subscales, subscale length, and subscale reliability. The largest gains are seen in cases where the correlations between subscales are high, the number of items on the subscale being augmented is low, and the reliability of the subscale providing ancillary information is high.

Stone et al. (2010) compares Yen's (1987) OPI; the regression-based approach described by Wainer et al. (2001); and a MIRT approach. They found that all three approaches to augmenting subscale scores that were examined improved the precision of the subscale scores

markedly. Although all three methods increased the precision of subscale scores, there are practical issues that differentiate the methods. For example, the MIRT modeling approach is significantly more complicated to implement whereas Yen's and the regression-based approaches are more easily implemented. In addition, higher inter-factor correlations were found for the MIRT approach indicating that less unique variance existed among the subscale scores. Another aspect to the methods is related to the scores that are produced and the comparability/interpretation of scores. Since the augmented subscores were more highly correlated with each other than non-augmented subscores, they were less able to differentiate between individual examinees' strengths and weaknesses.

Skorupski (2008) compared several augmented vs. non-augmented methods of subscore reporting: CTT, IRT, and Bayesian IRT. He found that all subscores showed a dramatic and relatively stable decrease in overall variability as a result of augmentation; the reliability estimate for each augmentation method increased the reliability of subscores by at least 20%. The most substantial gain in reliability occurred for those subscales with the fewest items and the smallest reliability prior to augmentation.

Skorupski's study, along with other studies examining augmentation (e.g. Stone et al., 2010; de la Torre, Song, & Hong, 2011), illustrates the fact that the purpose of subscore reporting is critical to the determination of what method of computation should be used. As mentioned before, the issue with subscore augmentation is that augmented subscale scores are more highly correlated and less variable than non-augmented scores. Depending on the method of augmentation, the subscores may become either more similar to the overall group's pattern, or to the individual student's other subscores. Using MIRT for subscore reporting may resolve this issue of making subscores less distinct. For example, de la Torre and Patz (2005) suggest that using MIRT methods that take into consideration correlations between subscores are particularly well suited

when a score profile is needed to determine a student's specific areas of strength and weakness. Scores from such a profile are more accurate and precise, and they can be used more reliably in determining the best learning trajectory for each student. However, when the use of test scores has high-stakes implications (e.g., certification, competition), sufficiently reliable scores derived without the benefit of ancillary information would be more appropriate.

Factors Affecting Subscore Reliability

Based on the variables related to subscore reporting methods investigated in the studies mentioned above, one can outline a number of different factors that affect reliability and precision under the CTT and IRT frameworks. A CTT-based test's reliability is a property of the scores for a particular group of examinees; it is not reliable or unreliable in its own right. Like other sample-dependent measurements, the reliability measure is most useful when the examinee sample is similar to the examinee population for whom the test is being developed. To the extent that the sample differs in some unknown way from the population, the utility of the item statistics may be reduced (Crocker & Algina, 1986; Hambleton & Jones, 1993). Reliability coefficients are also dependent on the heterogeneity of the sample; the same test taken by a heterogeneous sample will have a higher value of ρ than when taken by a homogeneous sample (Thissen, 2000). Crocker and Algina (1986) also mention that test speededness has an effect on reliability, in that variables other than ability may affect the test responses. Additionally, test length affects reliability, and longer tests are more reliable than shorter tests, which is evident from the Spearman-Brown formula and Cronbach's alpha formula. Increases in test reliability obtained from increasing test length follow the law of diminishing returns; in addition, adding items that are not consistent with other items on the test may make reliability calculated as

Cronbach's alpha lower. Additionally, the more people take the test, the more reliable the score estimate will be. Larger numbers of subjects make the statistics generated by that sample more representative of the population of people than would a smaller sample. These statistics are also more stable than those based on a small sample.

As mentioned previously, CTT measures of reliability are sample-specific; in addition, error is estimated for a group, rather than for individuals at specific score values. The item response analog of test score reliability and the SEM is the test information function (TIF). The use of TIF as a measure of accuracy of estimation is more appealing, because it is not sample-dependent, and it provides an estimate of the error of measurement at each ability level (Hambleton & Swaminathan, 1985). Since IRT reliability is dependent on the amount of information provided by the test, or the accuracy of estimating the person's location on the ability continuum, the factors that maximize information also minimize error of measurement. Specifically, the greater the slopes of the items that make up the test, the greater is the information provided by such items. Additionally, the maximum information is attained at the point where the person has a .50 probability of answering the item correctly. Items that are most informative also should have a low c parameter (guessing). As the c parameter decreases, the information increases, with the maximum information obtained where c=0. Test length has a positive impact on information: test information is considerably increased when test length increases (Hambleton & Swaminathan, 1985). While the number of people who took the test does not specifically affect the reliability of the test, it does affect parameter estimation. Large samples of people are needed for correct parameter estimation at all levels of θ. Additionally, CTT-based SEM tends to be smaller at the extremes of the scores, and larger in the middle range of the scores. IRT-based CSEM, on the other hand, tends to be smaller in the middle of ability distribution, and larger at the extremes.

A number of studies investigated the impact of the number of items in a subscale in the CTT and IRT environments. For example, Sinharay (2010) suggests that in order to have subscores that have added value, subscores should be based on an adequate number of items (at least 20 or more). However, if multidimensional scoring is used, having fewer items within a scale can result in sufficient reliability (de la Torre, Song, & Hong, 2011; De la Torre & Patz, 2005; Boughton, Yao, & Lewis, 2006; Wang, Chen, & Cheng, 2004, Yao, 2010).

A small number of studies considered the impact of the proportion of polytomous items on the test (Yen, 1997; Shin, 2007). No conclusive findings were made about this specific characteristic, aside from the fact that constructed response items tend to have higher reliability. Additionally, a small number of studies considered the differential impact of the items that contributed only to a single subscale, or to more than one subscale (Boughton, Yao, & Lewis, 2006).

Another factor that has been explored in relation to reliability is the correlation between subscales (Luecht, 2003; Boughton, Yao, & Lewis, 2006; Cheng, Wang, & Ho, 2009; Wang, Cheng, & Chen, 2004; De la Torre & Patz, 2005; De la Torre & Song, 2009; De La Torre, Song, & Hong, 2011). Different results were obtained depending on whether subscore augmentation was used in subscore reporting. Specifically, the subscore reporting methods that did not use subscore augmentation did not benefit from high correlations between domains. However, multidimensional scoring is found to be advantageous when there are several short subtests measuring highly correlated abilities that are also highly correlated with the ancillary information sources (de la Torre, 2009; Wang, Chen, & Cheng, 2004, Yao, 2010). It appears that for CTT/ non-augmentation methods, too high of correlation between domains may be a problem; reliability decreases if the correlations are too high. On the other hand, those methods that

borrow information from other domains/ MIRT methods, perform better when domains are highly correlated.

In language proficiency assessment, correlations between language domains change based on the students' age and level of proficiency, in addition to other factors. One reason for this change is that language proficiency does not develop uniformly in all domains. Proficiency development in one domain interacts with the proficiency development in all other domains; at one stage of proficiency development, these proficiencies in different domains interact with each other in different ways than at other stages of proficiency development. While the above finding has been firmly established (Solano-Flores & Trumbull, 2003; Chapelle et al., 2011; Alderson, 2005, 2007), the exact degree and methods of impact of proficiency in one domain on proficiency in others remains unclear (Chiappe & Siegel, 2006; Farnia & Geva, 2013). Numerous studies of some of those inter-domain interactions have been conducted. However, few studies examine the interaction between all four language domains, and those that do, have college students as subjects and are not longitudinal (Powers, 2010, 2013; Liu & Costanzo, 2013; Sawaki et al., 2008; Stricker et al., 2005). While many studies on the development of language abilities in L1 speakers exist, longitudinal studies of ELL proficiency development are relatively rare (Kieffer & Vukovic, 2013; Geva, 2000; August & Shanahan, 2006), and few extend beyond grade 2; therefore it is hard to say from the longitudinal perspective how the correlation between different domains changes across grades (Kieffer & Vukovic, 2013). Those studies that do exist concentrate on the development of the reading domain and various subskills thought to predict reading ability (Kieffer, 2011, 2012; Mancilla-Martinez et al., 2011; Mancilla-Martinez & Lesaux, 2010; Nakamoto et al., 2007). Out of those, even fewer deal with how the importance, impact, and interrelationships between different subskills change over the course of several years (Farnia & Geva, 2013; Ford et al., 2013; Manis, Lindsey, & Bailey, 2004).

The direction of the relationship between the domains is often unclear, and may change with age. For example, while it is widely believed that vocabulary and oral skills influence reading ability (Gottardo, 2002; Geva, 2006), some studies have shown that, on the other hand, it is the print exposure that explains significant variance in children's growth in vocabulary, verbal fluency, and general knowledge (Stanovich, 1993; Chiappe & Siegel, 2006). Similarly, the impact of grammar knowledge on reading ability is not consistently identified. For example, whereas grammatical knowledge appears to be strongly predictive of later reading for native English speakers (Lonigan, Schatschneider, & Westberg, 2008), a few studies have demonstrated a weaker relationship between grammatical knowledge and reading for ELLs in the elementary grades (e.g., Jongejan, Verhoeven, & Siegel, 2007; Lipka & Siegel, 2007).

At the same time that developmental views of reading support the cumulative nature of reading growth, they also highlight discontinuities in development that can result from the differential skills that play into reading during different ''stages'' or ''phases'' of reading development (e.g., Chall, 1983; RAND Reading Study Group, 2002). That is, the skills and knowledge that are required for comprehending fifth- or eighth-grade–level text are more sophisticated and varied than those required to comprehend a third-grade text, which in turn are greater and more varied than those required to read a first-grade text. Whereas variation in students' word reading skills generally explain the majority of variation in reading achievement in kindergarten through third grade, oral language proficiency and higher-order comprehension processes predict greater shares of the variance in reading achievement in Grades 4 and beyond (e.g., Catts, Hogan, & Adlof, 2005; Vellutino, Scanlon, Small, & Tansman, 1991).

While speaking and its components, such as vocabulary knowledge, have an undisputed impact on reading proficiency, it is possible that early oral English proficiency may become a weaker predictor of ELLs' reading achievement over time. For ELLs who enter school with

limited oral English, general ESOL instruction provides education in both oral language and reading. Their reading skills at this level are largely influenced by their knowledge of the language. By the middle school years, however, if they achieved general language proficiency comparable to that of native speakers, their reading achievement becomes less dependent on the knowledge of language rules and more comparable to their native English speaking classmates (Kieffer, 2011). Similarly, Farnia and Geva (2013) emphasize that the nature of the predictors and of reading comprehension changes over time and, therefore, that what predicts reading comprehension is not static. Specifically, longitudinal studies involving monolingual children have shown that performance in the early school years on word-level reading skills predicts a substantial amount of variance in later reading comprehension. However, in subsequent years, as word-level reading skills become established, language skills become stronger and more reliable predictors of reading comprehension. This general observation is supported in the studies of reading development in both native speakers and ELL students. Moreover, because language comprehension is complex and multidimensional, what constitutes language comprehension changes over time; therefore, different aspects of language comprehension predict reading comprehension. Relatedly, Gottardo and Mueller (2009) have shown that in the early school years, although oral language skills play some role in reading comprehension, the comprehension of texts relies extensively on word-level reading skills. However, language skills become more prominent in reading comprehension when the basic principles of word-level reading have been more or less established and the texts students read present more demanding language skills (e.g., Geva & Farnia, 2012; Whitehurst & Lonigan, 2002).

Several theories of interaction between writing and reading across time exist. A study by Davis and Bryant (2006), for example, found that at the age of 7, the ability to read influenced the ability to read, but then the causality of this relationship was reversed by age 10. Forrester

(2013) notes that both spelling and single word reading ability correlated highly with phonemic awareness at the beginning and middle of first grade. At the end of first grade these correlations became smaller, though still significant, suggesting that as children start acquiring knowledge about orthographic patterns, they rely less on using simple letter-sound correspondences in their reading and spelling.

Shanahan and Lomax (1986, 1988) found that reading and writing influenced each other, suggesting a dynamic relation, with knowledge arising from either reading or writing but then being generalized or diffused to the other process. They also found that the specific patterns of relation changed, with word recognition–spelling connections being relatively more important (as proportion of the variance) early on, but more structural aspects of text knowledge coming into prominence with older or more proficient readers.

The unique contribution of some subskills may change as a function of age. For example, Farnia and Geva (2013) found that phonological short-term memory emerged as a positive predictor of reading abilities only in Grade 4. Students whose reading ability predictors were studied at the age of 6-10 did not show that there was a correlation between these two variables. These authors found that phonological awareness, naming speed and working memory were significant early predictors of reading comprehension in Grade 6. However, performance on these three variables in Grade 4 did not contribute to Grade 6 reading comprehension. These changes may be explained by the possibility that the relationship between phonological awareness, and naming speed and later performance on reading comprehension is mediated through word recognition skills. Farnia and Geva (2013) also found that correlations between predictors changed across grades: different Grade 1 and Grade 4 individual cognitive variables predicted rate of growth in reading comprehension: in Grade 1, naming speed was a significant predictor, whereas in Grade 4, phonological short-term memory was significant.

The predictive power of the measures of reading ability shifts gradually over time from phonological awareness to print-related skills (Manis, Lindsey, and Bailey (2004). Once children begin formal literacy instruction, skills related to written language may become more highly related to later reading achievement than are sound-related skills alone. While oral language proficiency is important, oral language development has a strong relationship to emergent literacy skills in preschool, but that relationship weakens significantly in kindergarten and only becomes important again in third grade, when reading comprehension becomes the focus of instruction (Ford et al., 2013).

Storch and Whitehurst (2002) found that there was a strong relationship between oral language and early literacy skills in preschool; however, in kindergarten through second grade, the effect of oral language was found to be mediated by skills such as phonological awareness and print knowledge. By third and fourth grade, oral language became important once again because of its direct relationship with reading comprehension. These findings are consistent with research with English language learners. Although the relationship between English language proficiency and early word reading is usually not found to be robust, multiple language components, including vocabulary skills, syntactical knowledge, and listening comprehension, have all been shown to be associated with reading comprehension among English language learners of elementary-school age through adulthood (Geva, 2006). Thus, although oral language skills are clearly foundational for children's early literacy learning, evidence suggests that language ability pays bigger dividends to reading later in the developmental process when the characteristics of texts place greater demands on the reader for comprehension (Lindsey, Manis, & Bailey, 2003; Storch & Whitehurst, 2002).

Although research has demonstrated that phonological awareness, alphabet knowledge, and orthographic knowledge are inter-related (Schatschneider, Fletcher, Francis, Carlson, &

Foorman, 2004; Warley, Landrum, Invernizzi, & Justice, 2005), the relative importance of skills within these domains tends to change once formal literacy instruction begins. As children start to develop a more sophisticated understanding of written language, knowledge of the writing system quickly replaces sound-related skills as the most accurate predictor of later reading achievement (Hammill, 2004; Scarborough, 1998). Research has shown, for example, that from fall to spring of the kindergarten year, print-related skills replace phonological awareness skills as more robust predictors of reading achievement one and two years later (Morris, Bloodgood, & Perney, 2003b; Warley et al., 2005).

According to Genessee et al. (2005), several studies provide evidence of a positive relation between English oral proficiency and English reading. This relation holds across Grades 1–9 and for several different measures of oral proficiency and several different standardized measures of reading achievement. The relation between English oral proficiency and English literacy seems to strengthen substantially across the grades, arguably because both are similarly influenced by schooling and both are indicative of academic success.

In studying reading development in native speakers, some researchers have suggested that early reading comprehension is more strongly related to word decoding than to listening comprehension in monolingual English speakers (Wagner, Torgesen, & Rashotte, 1997), and relations among variables change over time (Hoover & Gough, 1990), with word recognition having a higher impact on reading ability in second grade than in eighth grade, but with listening comprehension having a lower impact in second grade than in eighth grade. Fourth-grade performance fell in between second- and eighth-grade performance in terms of relationships among variables (Catts, Hogan, & Adlof, 2005).

Based on the findings of the studies above regarding varying correlations between language domains, it is likely that these changes will affect the estimation of subscore

reliabilities for student groups at different grade bands and proficiency levels for those subscore estimation methods that take these correlations into consideration.

In the state assessment of language proficiency that is used for the present study, the number of items stays approximately the same in the domains across grade bands; in the speaking domain, the number of items remains exactly the same across grade bands. The number of subscales stays the same across grade bands. The number of students, item parameters, and correlations between domains change between grade bands and between proficiency levels. Due to the fact that we are using real data, we are not able to manipulate one single variable, such as the number of students per grade band or the number of items per domain, to examine its impact on the reliability and precision of subscores. This factor prevents us from pinpointing the source of variance in reliability and precision of subscores across grade levels and proficiency levels, or quantifying the impact of those characteristics on reliability and precision. However, as we find and interpret the differences in reliability and precision of subscores, inferences can still be made about the causes of variance based on the information from previous studies.

The reason for selecting the four subscore reporting methods (CTT, UIRT, augmented IRT, and MIRT) has to do with a wide range of approaches this selection covers. It allows us to compare the subscores derived within different frameworks (CTT vs. IRT), the treatment of collateral information (UIRT vs. augmented IRT), and the treatment of correlations between domains (CTT and IRT vs. augmented IRT and MIRT). In addition, while both augmented IRT and MIRT make use of the relationship between the domains, they do it differently. While augmented IRT estimates the variance-covariance matrix in a step that is separate from ability estimation, MIRT can estimate the relationship between the domains directly and do it in one step with ability estimation. While MIRT is used less frequently in large scale score reporting (Shin, 2007), other methods continue to be widely used in the testing industry.

Summary

The literature review provided the discussion of the following key elements: the review of reliability and precision concept and measurement methods under the CTT and IRT framework; the review of current subscore reporting methods; and the summaries of studies done to compare the reliability of subscore reporting methods. In addition, we reviewed the factors that affect reliability and precision in CTT and IRT frameworks, as applied to subscore reporting. Of those several factors, we concentrated on how the differences between inter-domain correlations across grade bands and the differences between inter-domain correlations across proficiency levels on the reliability and precision of subscore estimation.

CHAPTER 3

METHODS

This chapter contains three major sections: (1) a description of the student sample and testing instrument used in the present study; (2) a description of four methods of subscore reporting (number-correct CTT based by-domain raw score; raw by domain IRT score; augmented IRT score; and Bayesian MIRT score); and (3) the description of analytical procedures employed to evaluate the reliability and precision of each subscore reporting method.

Participants

The participants for the study were students in state public schools who took the English language proficiency assessment in 2013; therefore, all these students were considered to be ELL students. Approximately 48,215 students took the test. Approximately 48% were female, and 52% were male. The majority of the students in the sample were Hispanic (81%), with other ethnicities being Native American, Asian, Black, White, multiethnic, and Pacific Islander. The majority of students claimed Spanish as their home language (81%); the next most frequent home languages were Vietnamese and Arabic. Overall, K-12 ELL students in the state come from homes where close to 70 different languages are spoken. While the information about the native language of this sample is not available to us, the following information was obtained on the state K-12 ELL population in 2006:

First Language	Frequency	Percent
Arabic	581	1.76
Chinese	348	1.05
Chuukese (Marshall Island/Micronesian)	24	0.07
Dinka	27	0.08
Farsi (Iranian)	25	0.08
French	54	0.16
German	497	1.50
Hmong	254	0.77
Khmer (Cambodian)	107	0.32
Korean	249	0.75
Lao	401	1.21
Native American	63	0.19
Philippine or Tagalog	104	0.31
Portuguese	20	0.06
Russian	122	0.37
Sign Language	2	0.01
Somali	62	0.19
Spanish	26741	80.95
Thai	13	0.04
Vietnamese	1060	3.21
Yugoslavian (Bosnian/Serb/Croatian)	6	0.02
Other	1668	5.05
Missing	607	1.84
Total	33035	100.00

Table 1. Native Languages of ELL Students in the state (2006).

Table 2 represents the by-grade distribution of ELL K-12 students who took the assessment in 2013. After cleaning the data and eliminating all duplicates and students for whom only some subscores were available, it was found that a total of 43,708 students who met these conditions took the assessment in 2013 (table 2).

grade	K	1	2	3	4-5	6-8	9-12	total
Number of students	3649	4429	4389	4458	8392	10363	8028	43708

Table 2. Number of Assessed ELL K-12 Students by Grade Band.

The State Board of Education adopted four performance level names to describe the quality of student achievement demonstrated on the assessment. Those performance levels, from lowest to highest, were labeled: (1) beginning; (2) intermediate; (3) advanced; and (4) fluent. An extended Modified Angoff procedure (Hambleton & Plake, 1995) was used to recommend cut scores for seven grade levels (K, 1, 2, 3, 4-5, 6-8, 9-12) in each of the four domains (Reading, Writing, Listening, and Speaking). Participants in each grade/domain area recommended three cut score locations (Beginning/Intermediate, Intermediate/Advanced, and Advanced/Fluent). Table 3 indicates the number of students in each grade band by proficiency level.

	Proficiency level				
	1	2	3	4	total
gr0	445 (12%)	1237 (34%)	1318 (36%)	649 (18%)	3649
gr1	433 (10%)	1911 (43%)	1442 (33%)	643 (15%)	4429
gr2	384 (9%)	1405 (32%)	1554 (35%)	1046 (24%)	4389
gr3	276 (6%)	904 (20%)	1543 (35%)	1735 (39%)	4458
gr45	411 (5%)	1864 (22%)	2940 (35%)	3177 (38%)	8392
gr68	311 (3%)	1790 (17%)	3955 (38%)	4307 (42%)	10363
gr912	305 (4%)	1125 (14%)	2083 (26%)	4515 (56%)	8028

Table 3. Number of Students by Total Score-based Proficiency Level.

Instrument

The data for this study were collected from the state English Language Proficiency Assessment for the year 2013. This assessment is administered once a year, at the end of the

school year. According to the assessment technical manual (Peyton et al., 2009), it is designed for these purposes:

(1) to measure specific indicators within the state Curricular Standards for English to Speakers of Other Languages (ESOL) to indicate a student's level of proficiency with the English language in reading, writing, listening and speaking;

(2) to produce data that capture trends across the state and can measure progress in meeting Annual Measurable Achievement Objectives (AMAOs) for Title III accountability requirements; and

(3) to provide data on which to base decisions about designing instruction for English Language Learners (ELLs).

Generally, any student identified as an English Language Learner (ELL) or Limited English Proficient (LEP) based on a prior year's administration of the state assessment, or according to one of the commercially available assessments approved by the state department of education, is required to take the assessment. In addition, a student whose home language is other than English and who may not have been assessed for English proficiency after enrolling in the district, needs to take the assessment. An ELL student may exit an English for Speakers of Other Languages (ESOL) program by achieving "fluent" performance level on all four domains (Reading, Writing, Listening, and Speaking) and the total composite score of the assessment for two consecutive years. However, if later on a student who exited the program is found to be not doing well in academic content areas due to language proficiency, the school district may choose to administer the assessment to determine need for reclassification.

The assessment was developed to measure the targeted learning outcomes provided in the state Curricular Standards for English to Speakers of Other Languages (ESOL). ESOL standards are aligned to the State's Reading and Writing Standards and linked to the language of the State's Science and Mathematics Standards. The Curricular Standards for English to Speakers of Other Languages serve as the basis for what is assessed by the test.

The assessment measures ELL performance in four domains – reading, writing, listening, and speaking. The assessment was designed to span 5 grade bands: K-1, 2-3, 4-5, 6-8, and 9-12 (table 3). The total time allowed for test administration was 2 hours. The testing window is open from the beginning of February to the beginning of May.

Item formats and student response mode vary across the assessment domains. The assessment is made at the domain level; no further subskill assessment is made. In the speaking domain assessment, the questions are individually administered to students regardless of grade level, students respond orally and the examiner scores the response to each question immediately after it is given, and then records the answer onto an answer sheet. The items sets in the reading, writing, and listening domain assessments for students in the second grade and higher are intended to be group administered when appropriate. However, individual student circumstances (language, development, or disability limitations) may make it necessary that these test domains be individually administered (Peyton et al., 2009).

The Speaking subscale consists of constructed response (CR) items that are administered individually to students. These polytomous items may have 3, 4, or 6 categories. Questions and prompts are presented orally to the students. Students respond orally to the items; the tasks include answering short questions; elaborating on a question; and describing what is happening on a picture or a picture sequence. Six of the nine items comprising the Speaking domain are

scored using a scoring rubric of 0 to 2 or 0 to 3. For this set of items, students are rated on the appropriateness of the response to the item in addition to the number of sentences used, the completeness and grammatical correctness of the sentences used. The additional sets of three items are scored on a scale of 0 to 5 and are based on the elaboration and detailed description of a response such as to storytelling or description of a prompt (appendix 2).

The Listening subscale consists of a mixture of multiple choice (MC) and CR items in K and 1st grade, with polytomous items having 2 or 3 categories, and of dichotomously scored MC items only in grades 2-12. Items are administered individually to K and 1st grade, and as a group for older students. Questions and response choices are presented orally to the students; written presentation of items is not permitted. A CD recording of items is provided for a more standardized administration. The tasks include following directions; identifying beginning, middle, and ending word sounds; ability to distinguish between a grammatically correct vs. incorrect sentence; and listening comprehension questions based on a story read to the students.

The Reading subscale consists of MC items that are administered individually to K and 1st grade, and as a group for older students. The students read the items from the booklets and circle the selected answer from a list of options. The tasks included identifying rhyming words; completing cloze sentences; identifying synonyms/antonyms; word definitions; distinguishing between fact and opinion; analogies; and reading comprehension of passages. Oral presentation of the reading items by a teacher is not permitted.

The Writing subscale for K-1st grade consists of MC items administered individually. The tasks included writing letters/numbers based on oral prompt; completing cloze sentences; correctly rewriting sentences with syntactic errors; identifying correctly spelled word; and vocabulary identification (writing the word to label a picture). For grades 2-12, the writing section was split 50-50 between a MC section (dichotomously scored) and a CR writing performance section

(polytomously scored). The MC section included the identification of grammatically correct use of parts of speech; identification of synonyms/antonyms; punctuation, and syntax.

For the Writing Rubric section of the Writing domain students were given 20 minutes to write an essay based on either a picture or a written prompt. Students' writing responses were scored locally by trained school or district teachers and staff using a 0 - 4 point scoring rubric based on five writing traits developed by the ELL advisory group (appendix 2, table 1). The five traits for which ratings were to be systematically evaluated are vocabulary, sentence fluency, grammar, mechanics, and organization and development of a topic.

Number of items	K	Points	Gr 1	Points	Gr 2	Points	Gr 3	Points	Gr 4-5	Points	Gr 6-8	Points	Gr 9-12	Points
Speaking	9	31	9	31	9	31	9	31	9	31	9	31	9	31
Listening	18	22	18	22	22	22	22	22	22	22	22	22	23	23
Reading	24	24	24	24	23	23	23	23	23	23	29	29	28	28
Writing	14	14	14	14	15	15	15	15	15	15	17	17	17	17
Writing rubric	-	-	-	-	5	20	5	20	5	20	5	20	5	20
Total	65	91	65	91	74	111	74	111	74	111	82	119	82	119

Table 4. Assessment Structure: Number of Items and Possible Points in Each Domain.

Three parallel forms were developed for each grade band; however, only two of those forms were administered (alternating between odd and even years), and the third form was repurposed to be used as a brief identification or placement tool for English Language Learners new to the country or state.

Current Method of Total Score and Subscore Reporting

Scaling and test score equating for the state English Language Proficiency Assessment entailed four steps: (1) scaling the assessment by fitting item response data to an IRT model, (2) equating item parameters across forms within grade levels, (3) creating expected true score functions for each domain by converting person ability levels (θ) into expected true scores on a percent-correct metric by means of the test characteristic curve (TCC), and (4) calculating total composite scores as a weighted combination of domain scores. Only two form sets (one form per grade band) were used for proficiency testing purposes; the form sets were alternated between odd and even years. For each grade level (K, 1, 2, 3, 4-5, 6-8, and 9-12) and form (three initial operational forms for each grade level) of the assessment, a separate GPCM calibration was performed using PARSCALE (Muraki & Bock, 1997). Once all item parameters within a grade level were estimated and placed onto a common scale, TCCs were created for each content domain to determine the transformation from ability level (θ) to expected true score (on the percent-correct metric). Those expected true scores were then reported as final test scores. First, domain scores were calculated for each student within the grade band; then domain scores were combined into total scores using the weights specific for each domain and grade band. Weighting and the value of the weights applied to each of the domains was based upon consideration from the ELL Advisory Board, the state department of education, and the organization that produced the test. For grades 2 and higher, the writing domain was divided into two sections: Writing-Multiple Choice and Writing-Performance, each of which was worth 50% of the writing domain score.

Grade	Domain			
	Speaking	Reading	Writing	Listening
K	35	15	15	35
1	30	20	20	30

2	25	25	25	25
3	15	30	30	25
4-5	15	30	30	25
6-8	10	30	30	30
9-12	10	30	30	30

Table 5. Weights (%) by Domain and Grade Level Applied to Calculating Total Scores (Peyton et al., 2009).

To assess the construct validity of the assessment, the statistical modeling program MPLUS 5.21 (Muthén & Muthén, 2008) was used to examine the underlying factor structure of the assessment. A series of confirmatory factor analyses using a robust weighted least squares estimator was conducted to investigate the underlying factor structure for each of the five grade band test forms. A higher-order factor model was hypothesized to best explain the covariation in the items and domains. Specifically, a higher-order general factor (representing the total composite score) and four first-order factors for reading, writing, listening, and speaking were examined and tested (Peyton et al., 2009).

Due to the fact that the test combines MC and CR items, the generalized partial credit model (GPCM, Muraki, 1992) was used to score the items. As mentioned previously, Muraki's 1992 GPCM is a 2PL IRT model expressing the probability of selecting a particular response category over the previous one. It can be expressed as follows:

$$p\left(x_{jk} \mid \theta, a_j, b_j, \tau\right) = \frac{\exp[\sum_{h=0}^{k} a_j(\theta - b_j + \tau_{jh})]}{\sum_{c=1}^{m_j} \exp[\sum_{h=1}^{c} a_j(\theta - b_j + \tau_{jh})]},$$

where m is the number of response categories, k=1, …, m, and $\tau_1 = 0$. The discrimination parameter a_j indicates the degree to which categorical responses vary among items as θ changes. τ_h may be interpreted as the relative difficulty of step h in comparing other steps within an item; this difficulty may also be interpreted as the difficulty of endorsing a particular category (De

Ayala, 2009). Similar to the PCM, the categories do not have to be in ascending order, as the initial categories may be more difficult to endorse than the subsequent ones. However, in the present assessment, the categories are in ascending order, based on the grading rubric.

Data Analysis

Different Methods of Subscore Estimation

Domain scores were obtained by four methods: raw CTT number-correct score; unidimensional IRT domain scores; augmented unidimensional IRT; and multidimensional IRT.

CTT (Number Correct) Subscore Estimation

To obtain subscores for CTT, we used the SUBSKOR program (Skorupski, 2008). SUBSKOR is a program designed for implementing various subscore augmentation procedures in the CTT and IRT environment. The program provides subscore estimates, augmented and non-augmented subscore sample means, sample standard deviation estimates, and sample variance estimates. Reliability data is computed for each set of subscores.

CTT Reliability Estimation

For CTT, we calculated reliability using Cronbach's α. The formula is as follows:

$$\alpha = \frac{k}{(k-1)}\left(1-\frac{\sum \sigma_{x_i}^2}{\sigma_x^2}\right),$$

where k = number of items, $\sigma_{x_i}^2$ is the item variance, *and* σ_x^2 is the person variance.

SEM was calculated as follows:

$SEM = \sigma_{obs}\sqrt{(1-\rho_{xx})}$.

For the purpose of comparing precision at specific score levels derived within the CTT and IRT frameworks, CSEM for the CTT scores was calculated as follows based on the method proposed by Lord (1955):

$$CSEm = \sqrt{[\frac{X(k-X)}{k-1}]},$$

where X is the integer observed score and k is the number of items. While this formula does not take into account the fact that forms may differ in content, item difficulty, and other characteristics, since in our situation the items in a subscale are already grouped by those characteristics, this factor may ameliorate the overestimation of CSEM.

UIRT Subscore Estimation

We calculated the UIRT scores using the GPCM formula described above for each individual domain, similarly to how the UIRT scores are being obtained presently. We used the SUBSCOR software to calculate the scores. We calculated IRT scores for every domain separately with this method. Since the item parameters are known to us, we calculated θ based on the item parameter information for listening, speaking, reading, and writing.

We used EAP estimates to calculate ability parameters. The following formula was used to calculate the subscores in the GPCM model:

$$p\left(x_{jk} \mid \theta, a_j, b_j, \tau\right) = \frac{\exp[\sum_{h=0}^{k} a_j(\theta - b_j + \tau_{jh})]}{\sum_{c=1}^{m_j} \exp[\sum_{h=1}^{c} a_j(\theta - b_j + \tau_{jh})]},$$

where m is the number of response categories, k=1, …, m, and $\tau_1 = 0$.

EAP estimate has been cited as favored over the MLE and MAP estimate, and is considered the method of choice by a number of authors (Wang et al., 2004; Mislevy &

Stocking, 1989). Bock and Mislevy (1982) show that the EAP method produces reasonably accurate estimates of the person parameters. The accuracy of the EAP sample standard errors has also been studied (DeAyala, Schafer, & Sava-Bolesta, 1995). The EAP estimate of an individual's θ after administering L items is

$$\hat{\theta}_i = \frac{\sum_{r=1}^{R} X_r \, L\,(X_r) A\,(X_r)}{\sum_{r=1}^{R} L\,(X_r) A\,(X_r)},$$

where R is the number of quadrature points; $X_r$is the midpoint of a quadrature interval known as the quadrature node or point and has an associated weight $A\ (X_r)$ which reflects the height of the function g ($\theta \mid \vartheta$) around X_r, and $L\ (X_r)$ is the likelihood function at X_r given a certain response pattern x=x_1 , …, x_L and a specific IRT model.

UIRT Reliability Estimation

Within the IRT framework, measurement errors are no longer homogeneous along the θ distribution; instead, they depend on the levels of the latent trait. Each item provides a different degree of measurement precision at each θ level, which is referred to as the item information, denoted as I (θ), and test information is information summed across all items:

$$\mathrm{T}\ (\theta) = \sum_{i=1}^{n} \frac{P_i'^2}{PiQi}\ .$$

The test reliability within IRT is no longer a constant, but rather is a function of θ. The error variance at a given θ level, known as SEM, is inversely related to information and is calculated as follows:

$$\mathrm{SEM} = \frac{1}{\sqrt{I(\theta)}}.$$

CSEM is already calculated in IRT, since SEM is calculated for specific levels of θ. Specifically, since we used MML (marginal maximum likelihood) and EAP subscore estimation, we also used EAP reliability estimates. The standard error of measurement for EAP, which is the standard deviation of the posterior, is estimated as follows:

$$PSD\left(\hat{\theta}\right) = \sqrt{\frac{\sum_{r=1}^{R}(X_r - \hat{\theta}_i)^2 \, L\,(X_r) A\,(X_r)}{\sum_{r=1}^{R} L\,(X_r) A\,(X_r)}} \quad ,$$

where R is the number of quadrature points; $X_r$is the midpoint of a quadrature interval known as the quadrature node or point and has an associated weight $A\,(X_r)$ which reflects the height of the function g ($\theta \mid \vartheta$) around X_r, and $L\,(X_r)$ is the likelihood function at X_r given a certain response pattern $x=x_1, \ldots, x_L$ and a particular IRT model, such as a 1 PL, 2 PL, or a 3 PL model. The use of $PSD\left(\hat{\theta}\right)$ as the standard error is based on the fact that after 20 items the likelihood function and the posterior distribution are nearly identical, and the $PSD\left(\hat{\theta}\right)$ is pretty much interchangeable with the standard error (Bock & Mislevy, 1982; De Ayala, 2009). Usually the SEM calculated with the above methods is the value provided by IRT computer software.

Several different methods have been proposed to be able to compare the reliability of a set of scores obtained by several IRT methods (c.f. Mislevy et al., 1992; Wang, Chen, & Cheng, 2004; Wainer et al., 2001). Since we used Wainer et al.'s (2001) method of subscore augmentation, we also used the method of reliability estimation for non-augmented IRT used in this study. We used the following formula to do this:

$$\rho_v = \frac{Variance\,[EAP[\theta_v]]}{Variance\,[EAP[\theta_v]] + Average\,[SE^2[\theta_v]]},$$

in which $Variance\ [EAP[\theta_v]]$ is the variance of the IRT scale scores for subscale v, and $Average\ [SE^2[\theta_v]]$ is the average value of the variance of the error of measurement associated with those scores.

Augmented UIRT Subscore Estimation

We took the UIRT scores obtained previously without augmentation, and applied augmentation to them. We used SUBSKOR software to calculate the augmented UIRT scores. When subscores are estimated within an IRT framework, while the principle of subscore augmentation as borrowing information from elsewhere is retained, the implementation is slightly different. First, IRT subscores were estimated using Bayesian estimation methods (for consistency purposes, EAP estimation was used). The variable standard errors associated with different values of EAP(θ) was ignored, and they were treated as if the error of measurement was constant. A correction was made to the values of EAP(θ) to remove their regression towards the mean. After the values of EAP(θ) were computed, a formula analogous to Kelley's regression formula was applied:

$EAP\ [\theta_v] \approx\ \rho_v\ EAP*[\theta_v]$,

where ρ_v is an estimate of the reliability of subscale v, and $EAP*[\theta_v]$ is a hypothetical IRT scale estimate of θ_v that is not regressed toward the mean; that is, the values of $EAP*[\theta_v]$ are like the observed summed scores (Wainer et al., 2001).

Therefore to calculate the value of $EAP*[\theta_v]$,

$$EAP*[\theta_v] = \frac{EAP\ [\theta_v]}{\rho_v}.$$

To compute $EAP * [\theta_v]$ for each subscale, we needed some estimate of reliability of the IRT scale score for that subscale, ρ_v. This estimate of reliability was obtained based on the non-augmented IRT values as mentioned previously:

$$\rho_v = \frac{Variance\ [EAP[\theta_v]]}{Variance\ [EAP[\theta_v]] + Average\ [SE^2[\theta_v]]},$$

in which $Variance\ [EAP[\theta_v]]$ was the variance of the IRT scale scores for subscale v, and $Average\ [SE^2[\theta_v]]$ was the average value of the variance of the error of measurement associated with those scores. These values were computed as the usual summary statistics.

After computing an estimate of ρ_v for each subscale using the equation above, we computed $EAP * [\theta_v]$ for each examinee for each subscale, and then computed S^{obs} as the covariance matrix between the scale scores $EAP * [\theta_v]$. Once we have estimated S^{obs} and ρ_v, we computed S^{true} by subtracting from S^{obs} the diagonal matrix D, the diagonal elements of which were estimates of error variance. We used the equation

$S^{true} = S^{obs} - D$

to compute S^{true}, and then computed the empirical Bayes estimates of the IRT scale scores as follows:

$$\widehat{\underline{EAP}}[\theta] = \underline{EAP}[\theta]. + S^{true}\ (S^{obs})^{-1}(\underline{EAP}*[\theta]_j - \underline{EAP}[\theta].) =$$

$$\underline{EAP}[\theta]. + B * (\underline{EAP} * [\theta]_j - \underline{EAP}[\theta].)$$

where $\underline{EAP} * [\theta]_j$ was the vector of IRT scale scores for the subscales for examinee j; $\underline{EAP}[\theta].$ was the average vector of IRT subscale scores; $\widehat{\underline{EAP}}[\theta]$ was the vector of empirical Bayes estimates of the IRT scale scores, and

$B* = S^{true}\,(S^{obs})^{-1}$.

The weights in the matrix B* were the coefficients for the values of E$AP*[\theta_v]$ for each examinee. Because E$AP*[\theta_v]$ was computed from E$AP[\theta_v]$ by dividing each E$AP[\theta_v]$ by the corresponding subscale reliability estimate, we similarly obtained weights B by dividing B* by the reliability of the corresponding subscale:

$B = \frac{B*_{vv\prime}}{\rho_{v\prime}}$.

Then the weights B were used to compute the augmented estimates from the original values of $\underline{EAP}[\theta]$:

$\widehat{\underline{EAP}}[\theta] = \underline{EAP}[\theta]. + B(\underline{EAP}[\theta]_j - \underline{EAP}[\theta].)$

Then the weights from the B matrix were applied to the calculated subscore estimates to come up with the estimates of true subscore values for all four subscales:

$\hat{\tau}_{listen} = \beta_{0\,listen} + \beta_{listen}x_{listen} + \beta_{speak}x_{speak} + \beta_{read}x_{read} + \beta_{write}x_{write}$;

$\hat{\tau}_{speak} = \beta_{0\,speak} + \beta_{speak}x_{speak} + \beta_{listen}x_{listen} + \beta_{read}x_{read} + \beta_{write}x_{write}$;

$\hat{\tau}_{read} = \beta_{0\,read} + \beta_{read}x_{read} + \beta_{speak}x_{speak} + \beta_{listen}x_{listen} + \beta_{write}x_{write}$;

$\hat{\tau}_{write} = \beta_{0\,write} + \beta_{write}x_{write} + \beta_{speak}x_{speak} + \beta_{read}x_{sread} + \beta_{listen}x_{listen}$.

Augmented UIRT Reliability Estimation

We estimated augmented UIRT reliability using Wainer et al.'s (2001) method by calculating the ratio of the unconditional true score variance to unconditional estimated true score variance. The numerator of this fraction was the unconditional true score variance of the *v*th subscale, which was the *v*th diagonal element of the matrix

$A = S^{true}\,(S^{obs})^{-1}S^{true}\,(S^{obs})^{-1}S^{true}$.

The denominator, which was the unconditional score variance of the estimates for the *v*th scale, is the *v*th element of the matrix

$C = S^{true}\,(S^{obs})^{-1}S^{true}$.

Therefore the reliability of an augmented subscale could be expressed as follows:

$r_v^2 = a_{vv}/\,c_{vv}$.

The value of SEM was then calculated with the formula analogous to the SEM for CTT-derived scores:

SEM = $\sqrt{Variance\,[EAP[\theta_v]]} * \sqrt{1 - r_v}$.

Since Lord's formula for CSEM calculation for CTT scores is based on the number of points possible, there are no suggestions that it can be applied in the IRT environment. Only the average SEM is available for the augmented IRT scores.

MIRT Subscore Estimation

We used the generalized two-parameter partial credit model (M-2PPC) described by Yao and Schwarz (2006) as the multidimensional version of the GPCM, which is designed to describe the interaction of persons with items that are scored with more than two categories. The maximum score for item *i* is represented by K_i. To be consistent with the way dichotomous items are scored, the lowest score is assumed to be 0 and there are $K_i + 1$ score categories overall. The score assigned to a person on the item is represented by k = 0,1, …, K_i. Mathematically the MPCG model can be expressed as follows:

$$P\left(u_j = k\middle|\theta_j\right) = \frac{e^{ka_i\theta'_j - \sum_{u=1}^{k}\beta_{iu}}}{\sum_{\vartheta=0}^{K_i} e^{\vartheta a_i\theta'_j - \sum_{u=0}^{\vartheta}\beta_{iu}}}$$

where *k* is a score of person *j* in the category *u*, given the person's θ; a_i is the item discrimination parameter; β_{iu} is the threshold parameter of item *i* for score category *u*, β_{i0} is defined to be 0, ϑ is the number of steps in an item; and K_i is the highest score that can be obtained on item *i*.

To determine the point where the probabilities of obtaining the adjacent scores are equal, one needs to solve the equation

$$ka_i\theta_j' - \sum_{u=1}^{k}\beta_{iu} = (k+1)a_i\theta_j' - \sum_{u=0}^{k+1}\beta_{iu}.$$

The item information can be expressed by the following formula:

$$I_\alpha(\theta) = \frac{[\nabla_\alpha P(\theta)]^2}{P(\theta)Q(\theta)},$$

where α is the vector of angles with the coordinate axes that defines the direction taken from the θ -point, ∇_α is the directional derivative or gradient, in the direction α. Given the distance to the point and the directions from the axes, the values of the coordinates of the point on each of the axes can be recovered using the trigonometric relationship

$$\theta_\vartheta = \gamma \cos \alpha_\vartheta,$$

where θ_ϑ is the coordinate of the point on dimension ϑ, γ is the distance from the origin to the point, and α_ϑ is the angle between the ϑ th axis and the line from the origin to the point. The nominal categories model, of which a GPCM model is one representation, can be expressed in the form

$$z_k = a_k\theta + c_k,$$

in which z_k is a response value for category *k*, which is a linear function of θ with slope parameter a_k and intercept c_k (Thissen, Cai, & Bock, 2010). Another way to look at this model is by defining a set of unobservable variables, *v*, that follow the multiple common factor model. Thus, item *i* can be expressed as

$$v_i = \lambda' f + \delta_i,$$

where $\lambda' = [\lambda_1, \lambda_2, \ldots, \lambda_n]$ is the matrix of common factor loadings; *f* is a vector of common factors, and δ_i is the *i*th unique factor. The model assumes that for each item *i* there is a latent ability v_i that is required to correctly answer the item. This latent ability is assumed to be continuous and normally distributed. If the examinee's proficiency is beyond a given threshold, t_i, they will get the item correct. Therefore, for each dichotomous item i, the probability of an examinee answering correctly is

$$P = \{U_i = 1 \mid \theta_1 \ldots \theta_k\} = N\{\beta_{i0} + \beta'\theta_i\} = N\{\beta_{i0} + \beta_{i1}\theta_1 + \beta_{i2}\theta_2 + \ldots + \beta_{ik}\theta_k\},$$

where N is the normal ogive function, and the intercept $\beta_{i0} = \frac{t_i}{\sqrt{\psi_i}}$,

and for the *k*th dimension $\beta_i = \frac{\lambda_i}{\sqrt{\psi_i}}$,

where ψ_i is the explained item variance or 1 minus the communality, given as $\psi_i = 1 - \lambda_i' P \lambda_i$, where P is the covariance matrix of latent abilities. This formula is the extension of Lord's (1980) parameterization of the unidimensional model

$$P = \{U_i = 1 \mid \theta\} = N\{a_i(\theta - b_i)\} = N\{a_i\theta - a_i b_i\}.$$

Comparing the two models for the unidimensional case, β_{i0} is analogous to -$a_i b_i$, and thus cannot be interpreted as simply the difficulty or location parameter, b_i. β_i corresponds to the

discrimination parameter a_i (Ackerman, Gierl, & Walker, 2003). In the GPCM model, the step parameter d is added to the $(\theta - b_i)$.

In this study, multidimensionality was specified, rather than determined from the data, so the multidimensional approach is confirmatory, rather than exploratory. We based the assumption that the data is multidimensional on the factor analysis study done during piloting the assessment. We considered the domains to be distinct traits, and therefore the four separate dimensions of the test; thus each competency contributed some amount of unique variance to the overall proficiency trait. While several different MIRT models can be fitted to the data (e.g. bifactor model, two-factor model, a higher order IRT model etc.), and the fit can be compared to see which model fits the data best, our main goal was not to compare the fit of various models. Instead, our research was guided by the previously determined test structure as having four domains. We ran a confirmatory IRT model in which the number of dimensions was specified, but the correlations between the dimensions were unconstrained. That is, the latent trait correlations were estimated by the model. In addition, we used a compensatory model, in which a stronger ability compensated for a weaker ability.

Confirmatory models allow researchers to formulate a number of theories about dimensionality and item weights based on the researcher's theoretical knowledge and previous research findings. As the orthogonal factor structure (i.e., factors are constrained to be independent) is imposed in exploratory factor analysis for model identification, the dimensions in multidimensional item response models are usually constrained to be orthogonal. Under this constraint, the derived dimensions are independent, but the interpretation of factor meaning under this condition may not be straightforward. Once the dimensions are constrained to be independent, no collateral information from the interdimension correlations is available, and

therefore the multidimensional approach outlined in this study is no longer useful. Only if the orthogonality constraint is released will the multidimensional approach provide better measurement precision. To release this constraint, one should specify item loadings on specific dimensions; in this situation, the factor structure is specified rather than discovered from the data (Wang, Chen, and Cheng, 2004).

MIRT Subscore Estimation Software

We used flexMIRT (Houts & Cai, 2013) to calculate the subscores under a MIRT framework. FlexMIRT fits a variety of unidimensional and multidimensional IRT models to single-level and multilevel data using maximum marginal likelihood (MML) or modal Bayes via Bock-Aitkin EM (with generalized dimension reduction), or Metropolis-Hastings Robbins-Monro (MH-RM; Cai, 2010b, 2010c) estimation algorithm. FlexMIRT produces IRT scale scores using maximum likelihood (ML), maximum a posteriori (MAP), and expected a posteriori (EAP) estimation. It (optionally) produces summed-score to IRT scale score (EAP) conversion tables for unidimensional and multidimensional IRT models. FlexMIRT can estimate any arbitrary mixtures of 3-parameter logistic (3PL) model, logistic graded response model (which includes 2PL and 1PL as special cases), and the nominal categories model (including any of its restricted sub-models such as generalized partial credit model, partial credit model, and rating scale model) for both single-level and multilevel data, in any number of groups. It can both calibrate and score, or only score, or only calibrate the items depending on the user's needs (Houts & Cai, 2013).

According to Houts and Cai (2013), although Bock and Aitkin's (1981) EM algorithm made IRT parameter estimation practical, the method does have shortcomings, with the most noticeable one being its limited ability to generalize to truly high-dimensional models. This is

due primarily to the need to evaluate high-dimensional integrals in the likelihood function for the item parameters. As the number of dimensions of a model increases linearly, the number of quadrature points increases exponentially, making EM estimation unwieldy and computationally expensive for models with more than a handful of latent factors. MH-RM algorithm has been implemented in flexMIRT to allow for the estimation of higher dimensional models. Within the flexMIRT implementation, the MH-RM iterations use initial values found through an unweighted least squares factor extraction stage (Stage I) and further improved upon during a supplemented EM-like stage (Stage II) - both Stage I and Stage II use a fixed number of cycles. Stage III is where the primary estimation of parameters occurs. Cycles in Stage III terminate when the estimates stabilize.

Houts and Cai (2013) describe the Cycle j +1 of the MH-RM algorithm for multidimensional IRT as consisting of three steps:

1. Imputation. In the first steps, conditional on provisional item and latent density parameter estimates $\beta(j)$ from the previous cycle, random samples of the individual latent traits $\theta(j+1)$ are imputed using the MH sampler from a Markov chain having the posterior of the individual latent traits $\pi(\theta|Y,\beta(j))$ as the unique invariant distribution.

2. Approximation. In the second step, based on the imputed data, the complete data log-likelihood and its derivatives are evaluated so that the ascent directions for the item and latent density parameters can be determined later.

3. Robbins-Monro Update. In the third step, RM stochastic approximation filters are applied when updating the estimates of item and latent density parameters. First, the RM filter is applied to obtain a recursive stochastic approximation of the conditional expectation of the complete data information matrix. Next, the RM filter is used again when updating the new parameter

estimates. The sequence of parameters converges with probability 1 to a local maximum of the likelihood of the parameter estimates, given the observed data (Houts & Cai, 2013).

FlexMIRT parameterizes the GPCM model according to Thissen, Cai, and Bock's (2010) paper, where the a parameter is derived from the formula a=Tα with $\alpha_1 = 1, \text{and } \alpha_2, \ldots, \alpha_{m-1} = 0$, where m is the number of categories in a polytomous item, and c parameter is derived from the formula c=Tγ, where T is a contrast matrix and γ is a vector. Traditional a (discrimination) parameter can be derived from FlexMIRT a* parameter as follows:

$$a = \frac{a*}{1.7};$$

whereas traditional b (difficulty) parameter can be derived as $b_i = \frac{-\gamma_{i1}}{a_i^*}$.

Therefore $\gamma = -a * b$, which is analogous to the intercept β_{i0} in Ackerman, Gierl, and Walker's (2003) equation (see table XXX). The step d parameter can be calculated as

$$d_k = \frac{c_k - c_{k-1}}{a_i^*} - \frac{c_m}{a^* (m-1)}.$$

Essentially, the slope and intercept parameters are calculated from the factor loadings and the estimated covariance matrix between factors, and then using these values, the student θ values are calculated. To determine the step difficulty, the average b_i value for all categories is the overall b parameter for the item; the difference between the average b_i value and the step b value was the step difficulty. The value of the first step in a polytomous item is 0 and is not explicitly assigned in the output table.

MIRT Reliability Estimation

The idea of test information as reliability in MIRT is analogous to unidimensional IRT. As was the case for the unidimensional IRT models, the reliability of a test can be determined by summing the information available from each item. However, in MIRT models the sum of the information estimates must be for the same direction in the θ -space. The test information function at point $\vec{\theta}$ in θ space in the direction $\vec{\alpha}$ can be obtained by

$$I\,(\theta_\alpha) = \frac{1}{Var(\hat{\theta}_\alpha)}.$$

Consequently, similar to the UIRT reliability estimation,

$Var(\hat{\theta}_\alpha) = \frac{1}{I\,(\theta_\alpha)}$, and

$\text{SEM}\,(\theta_\alpha) = \sqrt{Var(\hat{\theta}_\alpha)}$.

CSEM in MIRT is calculated similarly to UIRT; it is essentially SEM at the specific level of θ.

To be able to compare the reliability of unidimensional and multidimensional IRT methods, we calculated the reliability in the same way as Wainer et al. (2001) suggested to calculate it for the non-augmented IRT. We used the following formula to do this:

$$\rho_v = \frac{Variance\,[EAP[\theta_v]]}{Variance\,[EAP[\theta_v]] + Average\,[SE^2[\theta_v]]},$$

in which $Variance\,[EAP[\theta_v]]$ is the variance of the IRT scale scores for subscale v, and $Average\,[SE^2[\theta_v]]$ is the average value of the variance of the error of measurement associated with those scores.

Method Comparison Criteria

To determine how the four methods of subscore reporting in the four domains compare in respect to reliability and precision across grade bands, some similarities and differences in the size of these statistics need to be identified. In the situations where the reliability and precision of the scores derived from simulated data are compared, one would simply compare the expected (true) and observed parameter values, or correlate them and evaluate the size of the RMSE (e.g. Yao & Boughton, 2007; Edwards & Vevea, 2006; De la Torre & Song 2009; De la Torre, Song, & Hong, 2011; De la Torre & Patz, 2005; Boughton, Yao, & Lewis, 2005). However, since we did not know the true value of subscores (i.e. the students' true ability) because we used real data, we needed to use alternative means of comparing the reliability and precision of subscores calculated by different methods. A number of authors who used real data for their studies report descriptive statistics of subscores obtained by different methods, such as mean, median, and standard deviation/ variance (Skorupski, 2008; Stone et al 2010; DeMars, 2005; Luecht, 2003) and correlations between subscores obtained by different methods (Skorupski, 2008; Wang, Chen & Cheng, 2004; Stone et al 2010; DeMars, 2005). Luecht (2003) evaluated subscores obtained by different computational methods within Campbell and Fiske's (1959) multi-trait, multi-method (MTMM) approach by creating a dot-density plot of the correlations between the subscores for each competency area and computational method.

In addition, the comparison between CTT and IRT-derived subscores, and unidimensional and multidimensional IRT-derived subscores may be less than straightforward, as their frameworks operate under different assumptions and have different concepts of reliability and precision. As mentioned previously, a number of researchers have attempted to reconcile the differences between reliability and precision measurement in CTT and IRT (e.g. Lord, 1980; Green et al., 1984; Dimitrov, 2002, 2003; Culpepper, 2013). Other researchers,

however (c.f. Samejima, 1977; Kim & Feldt, 2010), stress the incompatibility of CTT and IRT frameworks, pointing out that the CTT-based definition of reliability has little relevance for measurement based on IRT, where the error variance is expressed as a function of ability. Therefore in a situation where some comparison needs to be made, researchers may state that these comparisons cannot be interpreted directly (c.f. Wainer et al., 2001) and then include descriptive information of subscore mean, standard deviation and variance, and possibly graphs showing these values (c.f. Skorupski, 2008; DeMars, 2005; Stone et al., 2010; Wang, Chen, & Cheng, 2004).

Our first two research questions were as follows:

1. How do the four methods of subscore reporting for four language domains compare in terms of subscore reliability and subscore correlations across five grade bands?
2. How do the four methods of subscore reporting for four language domains compare in terms of subscore precision across five grade bands?

To answer them, based on the methods mentioned above, we reported the reliability and SEM/ CSEM for all four methods in the four language domains across five grade bands. Since another criterion for evaluating the various scoring methods relates to the measurement errors, we reported the mean and standard deviation/ variance of subscores derived by these different methods. We were thus able to compare the score variance of the four methods of estimation.

In addition, we reviewed subscore profiles obtained by these different computational methods for five randomly selected candidates in each grade band (cf. Luecht, 2003; Skorupski, 2008) and compared it to the average subscore profile for all examinees within the grade band (cf. Skorupski, 2008). This information, especially graphically exhibited, provides a good idea of the variability of subscores obtained by different methods.

We also performed correlations between subscores for each domain obtained by these different methods. In this way, we compared reliability and precision between domains for each grade band. With the multidimensional approach, the variance–covariance matrix is estimated directly, and these estimates are unbiased. However, when the unidimensional approach is applied, direct estimation of the variance–covariance matrix or the correlations between latent traits is not possible. To obtain the correlations between latent traits when the unidimensional approach is applied, first Pearson product-moment correlation between the subscores is calculated:

$$\rho_{XY} = \frac{\sum_{i=1}^{n}(X_i-\bar{X})(Y_i-\bar{Y})}{\sigma_X \sigma_Y},$$

where σ_X and σ_Y are standard deviations of X and Y, respectively.

We performed correlations of subscores for the same domain obtained by different computational methods to assess the degree of difference between the subscores (cf. Luecht, 2003). Additionally, we reported average examinee subscores of ability estimates obtained with these four different methods for all grade bands and domains. This method of subscore estimation comparison was used by Skorupski (2008). Currently, the percent correct (expected) true score equivalent is determined for the possible range of raw number-correct scores and the IRT score of a student. We determined the percent correct (expected) true score equivalents for augmented IRT and MIRT scores to perform this comparison.

Our third research question is as follows:

3. How does the precision of these four methods of subscore reporting in the four language domains compare in the estimation of ability of examinees within five grade bands at different proficiency levels?

To answer it, we calculated CSEM for the CTT, UIRT, and MIRT methods of subscore reporting at the integer values of θ for all grade bands and domains and compared them. This comparison allowed us to compare CSEM estimated for different methods of subsore reporting at different proficiency levels. The CSEM for augmented IRT cannot be calculated due to the fact that for this method of subscore estimation, only the average SEM is available.

While test score reliability has the advantage of being conveyed as one number, the distribution of test information that emphasizes conditional measurement is often inspected visually in a particular scale (Shu & Schwartz, 2014). Therefore we plotted the CSEM over the test information curve for the four language domains and five grade bands at the integer levels of ability parameters, including those that are cutoff points for the four levels of proficiency.

Summary

This chapter reviewed research methods that will be used in this study, including the four methods of subscore calculation, the methods of reliability estimation for the four methods, and how these methods will be compared based on their reliability. The results of these comparisons helped us answer the research questions regarding how the four methods of subscore reporting compare in their reliability, and how the reliability and precision of subscores obtained by the four methods differs between domains across different grade bands and ability levels.

CHAPTER 4

RESULTS

The focus of this chapter is on presenting the results of the analyses of ELL students' responses to the items of the 2013 state English language proficiency assessment in 4 domains – listening, reading, writing, and speaking – scored by four methods – number correct (CTT), unidimensional IRT, augmented IRT, and multidimensional IRT, and to compare the reliability and precision of the four methods of subscore assignment.

Students in seven grade bands were assessed; some grade bands consisted of just one grade (grade 0, or kindergarten; grade 1; grade 2; and grade 3); other grade bands consisted of more than one grade (grade bands 4-5, 6-8, and 9-12). Some of the domains, such as speaking and writing rubric, consisted of only polytomous items; other domains, such as reading and multiple choice (MC) writing, consisted of only dichotomous items. The listening domain in grades 0 and 1 consisted of both dichotomous and polytomous items, and from grade 2 on – from dichotomous items only. Therefore the number of items in a domain ranged from 9 to 29, and the number of possible points in a domain ranged from 14 to 31; the total number of items in the assessment ranged from 65 to 82, and the total possible score ranged from 91 to 119.

This study addressed the following questions:

2. How do the four methods of subscore reporting for four language domains compare in terms of subscore reliability and subscore correlations across five grade bands (K-1, 2-3, 4-5, 6-8, and 9-12)?
2. How do the four methods of subscore reporting for four language domains compare in terms of subscore precision across five grade bands?

4. How does the reliability and precision of these four methods of subscore reporting in the four language domains compare in the estimation of ability of examinees within five grade bands at different proficiency levels?

The results of the study were determined by computing the subscores for each of the four domains for all seven grade bands using four methods – number correct (CTT), unidimensional IRT, augmented unidimensional IRT, and multidimensional IRT, and calculating the subscore correlations, reliability, and error of measurement.

Descriptives

Descriptive information about subscores was obtained. The means within the same domain were increasing across all domains from grade 0 to grade band 9-12. On average the CTT mean for listening was 16.84; for reading – 17.80, for multiple choice (MC) writing – 11.02; for speaking – 24.50, and for writing rubric – 14.04. The means of IRT-derived scores for all domains and grade bands were close to 0, which is the mean of normally distributed ability. On average the UIRT and augmented IRT mean for listening was .70, for reading - .74; for MC writing - .78; for speaking – 1.18; and for writing rubric - .88. For MIRT, the mean for listening was .20; for reading - .21; for MC writing - .21; for speaking - .18; and for writing rubric - .27.

Generally, the more points were possible in a domain, the higher the SD was for that domain. The standard deviation of CTT-derived subscores was the highest of all four methods. Therefore, the highest CTT subscore SD on average was for speaking (4.8 for 31 points possible), with the SD for other domains being somewhat lower (the average SD for listening was 3.01; for reading – 4.67; for MC writing – 2.66, and for writing rubric – 3.37). The average SD for UIRT subscores for listening was 1.32; for reading – 1.27; for MC writing – 1.21; for speaking – 1.46; for writing rubric – 1.42. Augmented IRT subscore SD was consistently lower

than the UIRT SD. The average SD for augmented IRT subscores for listening was 0.94; for reading – 1.10; for MC writing – .96; for speaking – 1.30; and for writing rubric – 1.28. MIRT subscore SD was consistently lower than the UIRT subscore SD, and was also almost always lower than the augmented IRT subscore SD. The only two instances where the MIRT subscore SD was higher than the augmented IRT subscore SD was the listening and MC writing domains of grade 0. EAP method of estimation, which tends to reduce the variance of the estimated scores, was used for IRT-based methods. The average SD for MIRT subscores for listening was .81; for reading – .86; for MC writing – .84; for speaking – .78; and for writing rubric - .80. Tables 6-9 provide examples of means and SDs for all grade bands by subscore estimation methods for all domains.

CTT	Gr0	Gr1	Gr2	Gr3	Gr45	Gr68	Gr912
	13.57	16.38	16.49	18.02	17.08	18.08	18.29
L	(3.30)	(2.99)	(3.33)	(2.91)	(2.27)	(2.98)	(3.27)
	11.31	20.0	12.73	16.33	17.63	23.95	22.66
R	(5.27)	(4.29)	(4.93)	(4.78)	(4.23)	(4.47)	(4.72)
	6.41	9.56	9.49	11.4	12.44	13.38	14.46
W	(2.21)	(2.48)	(3.35)	(3.00)	(2.27)	(2.66)	(2.68)
	17.01	21.52	24.18	25.53	26.75	28.11	28.39
S	(6.63)	(5.54)	(4.87)	(4.42)	(4.10)	(3.78)	(4.31)
			11.71	12.82	14.22	15.45	16.02
WR	-	-	(3.48)	(3.39)	(3.41)	(3.21)	(3.36)

Table 6. Means and SDs (in parentheses) for All Grade Bands for Subscores Estimated for the Four Domains with the CTT Method.

UIRT	Gr0	Gr1	Gr2	Gr3	Gr45	Gr68	Gr912
	.17	.30	.48	.87	.93	1.04	1.12
L	(1.30)	(1.26)	(1.41)	(1.34)	(1.35)	(1.27)	(1.29)
	.29	.71	.36	.77	.86	1.07	1.09
R	(1.14)	(1.36)	(1.34)	(1.27)	(1.27)	(1.17)	(1.32)
	.28	.39	.40	1.64	.79	.89	1.08
W	(1.09)	(1.23)	(1.34)	(1.13)	(1.26)	(1.17)	(1.23)
	.67	.74	.93	1.20	1.27	1.71	1.75
S	(1.46)	(1.32)	(1.46)	(1.53)	(1.55)	(1.48)	(1.45)
			.39	.55	.86	1.15	1.46
WR	-	-	(1.46)	(1.36)	(1.49)	(1.36)	(1.42)

Table 7. Means and SDs (in parentheses) for All Grade Bands for Subscores Estimated for the Four Domains with the UIRT Method.

A- IRT	Gr0	Gr1	Gr2	Gr3	Gr45	Gr68	Gr912
	.17	.30	.48	.87	.93	1.04	1.12
L	(.87)	(.87)	(1.03)	(.96)	(.94)	(.95)	(.98)
	.29	.71	.36	.77	.86	1.07	1.09
R	(1.01)	(1.20)	(1.15)	(1.10)	(1.06)	(1.00)	(1.15)
	.28	.39	.40	1.64	.79	.89	1.08
W	(.81)	(.95)	(1.10)	(1.05)	(.93)	(0.87)	(1.02)
	.67	.74	.93	1.20	1.27	1.71	1.75
S	(1.33)	(1.18)	(1.28)	(1.33)	(1.35)	(1.30)	(1.32)
WR	-	-	.39	.55	.86	1.15	1.46

(1.30)	(1.21)	(1.35)	(1.22)	(1.30)	

Table 8. Means and SDs (in parentheses) for All Grade Bands for Subscores Estimated for the Four Domains with the Augmented IRT Method.

MIRT	Gr0	Gr1	Gr2	Gr3	Gr45	Gr68	Gr912
	.01	.13	.10	.20	.24	.38	.35
L	(.89)	(.86)	(.89)	(.85)	(.77)	(.74)	(.63)
	.01	.14	.10	.21	.26	.39	.37
R	(.98)	(.91)	(.96)	(.92)	(.83)	(.77)	(.66)
	.01	.14	.11	.21	.25	.38	.36
W	(.93)	(.90)	(.95)	(.90)	(.80)	(.76)	(.63)
	.01	.14	.11	.21	.20	.34	.27
S	(.97)	(.94)	(.93)	(.90)	(.67)	(.59)	(.45)
			.11	.22	.27	.41	.36
WR	-	-	(.96)	(.93)	(.79)	(.74)	(.57)

Table 9. Means and SDs (in parentheses) for All Grade Bands for Subscores Estimated for the Four Domains with the MIRT Method.

Another illustration of variability reduction in augmented IRT and MIRT was the plots of subscores of five randomly selected students from each grade band. Augmented IRT subscores of any given individual were closer to the mean of the group subscores. On the other hand, MIRT subscores of an individual for a given domain were more like this individual' subscores for other domains. The highest variability between different domains for these five students was for the subscores obtained with the CTT and UIRT methods; augmented IRT and MIRT

variability was lower. This finding was consistent across all grade bands. Here we provided examples for grade 1 only for the four methods of subscore estimation (fig.12-15).

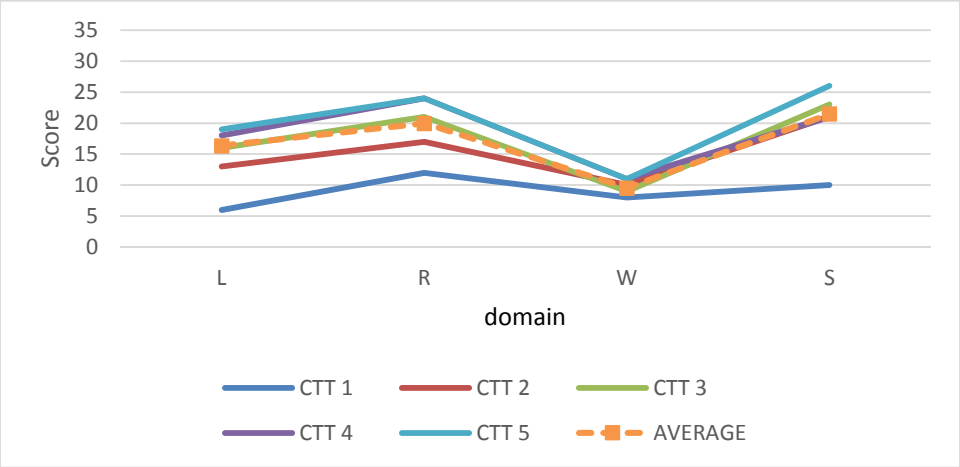

Figure 12. CTT subscore profiles in four domains estimated for five randomly selected grade 1 students, including an average subscore profile for all students.

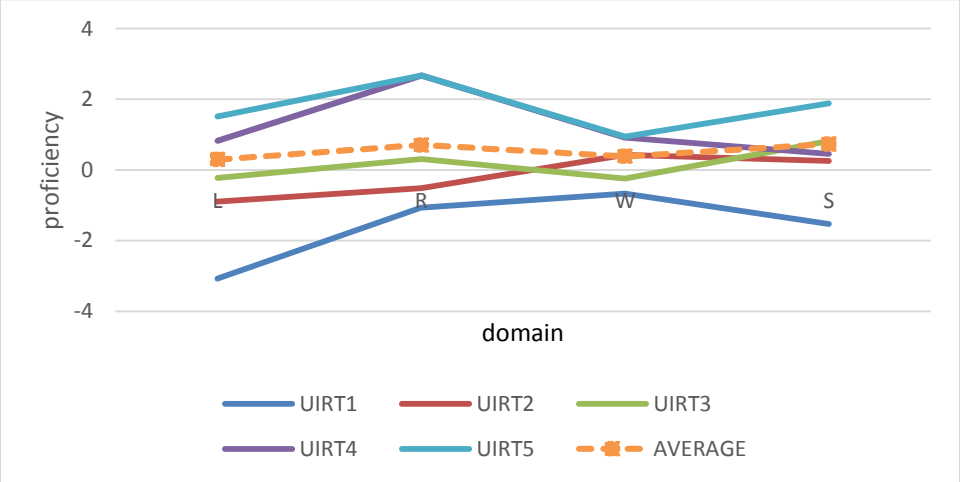

Figure 13. UIRT subscore profiles in four domains estimated for 5 randomly selected grade 1 students, including an average subscore profile for all students.

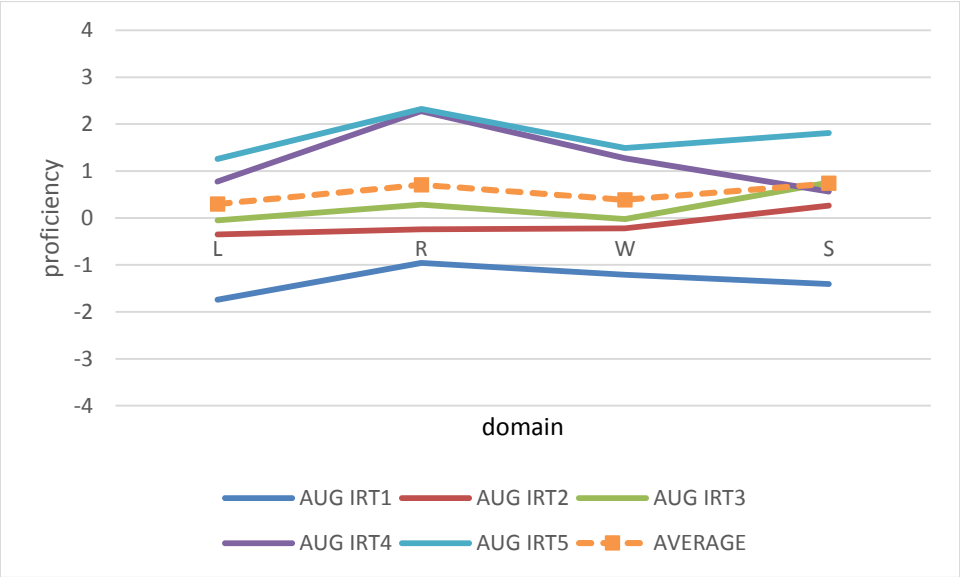

Figure 14. Augmented IRT subscore profiles in four domains estimated for five randomly selected grade 1 students, including an average subscore profile for all students.

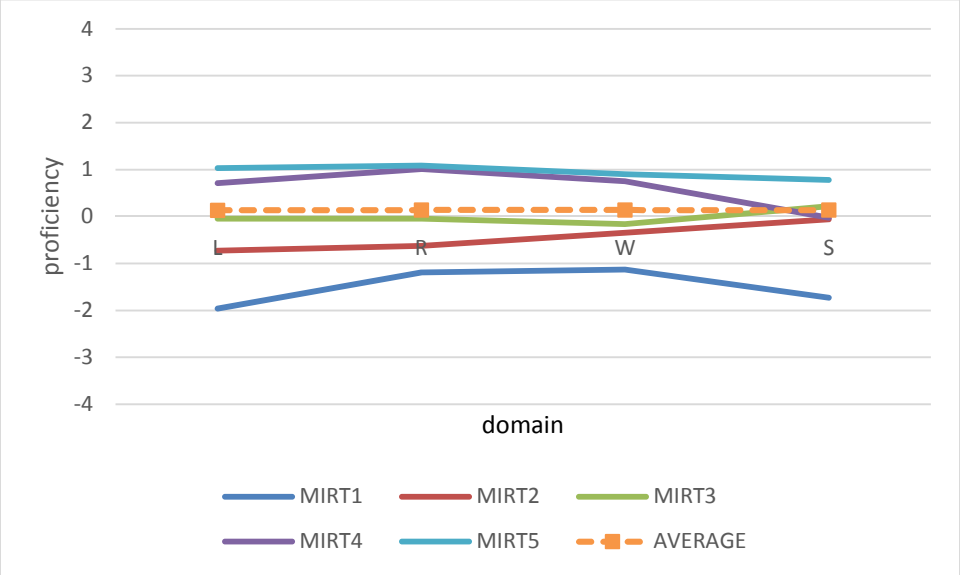

Figure 15. MIRT subscore profiles in four domains estimated for five randomly selected grade 1 students, including an average subscore profile for all students.

Another way to look at the variability of subscores is by constructing boxplots of the distribution of subscores obtained by different methods (figure 16). Here, again, we noted that

while the median was approximately the same for UIRT, augmented IRT, and MIRT-derived subscores, the variability of IRT-derived subscores was smaller than that of CTT-derived subscores. The variability of subscores derived by augmented IRT and MIRT methods was smaller than that of UIRT. At the same time, likely due to the decreased variability, there were more outliers in the subscores obtained by IRT methods, and specifically augmented IRT and MIRT methods.

When the data were examined for skewness and kurtosis, it appeared that the CTT-based subscores were fairly normally distributed in all four domains in grade 0, with only reading somewhat skewed to the right, indicating that there were more poor readers than good readers in that age group. For grade 2, however, all domains were skewed to the left, except for reading, which was distributed fairly evenly across possible scores. Otherwise, for all other grade bands the ability distribution was skewed to the left in all domains for the CTT-based subscores. For IRT-based scores, ability distribution was close to normal for grade bands 0, 1, 2, 3, and 4-5, and became more skewed to the left in grade bands 6-8 and 9-12. CTT-based subscores also exhibited the most skewness - ranging from -3.0 to .509, and kurtosis ranging from -.888 to 12.762, whereas IRT subscores exhibited skewness ranging from -1.224 to .322 and kurtosis ranging from -1 to 3.097. Figures 17-20 represent examples of subscore distribution for the speaking domain for grade 1 estimated with four different methods in the histogram form.

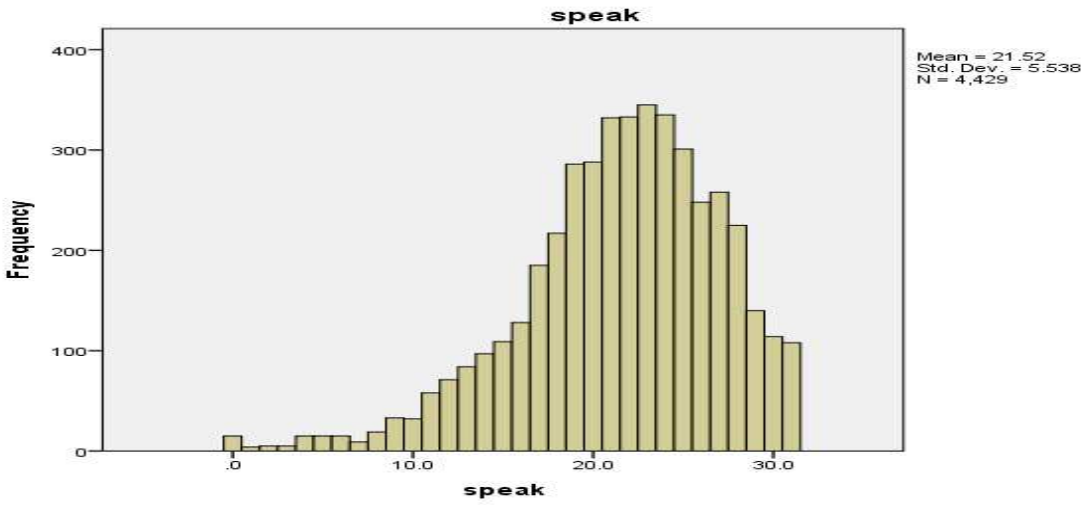

Figure 16. Speaking domain subscore distribution for grade 1 estimated with CTT.

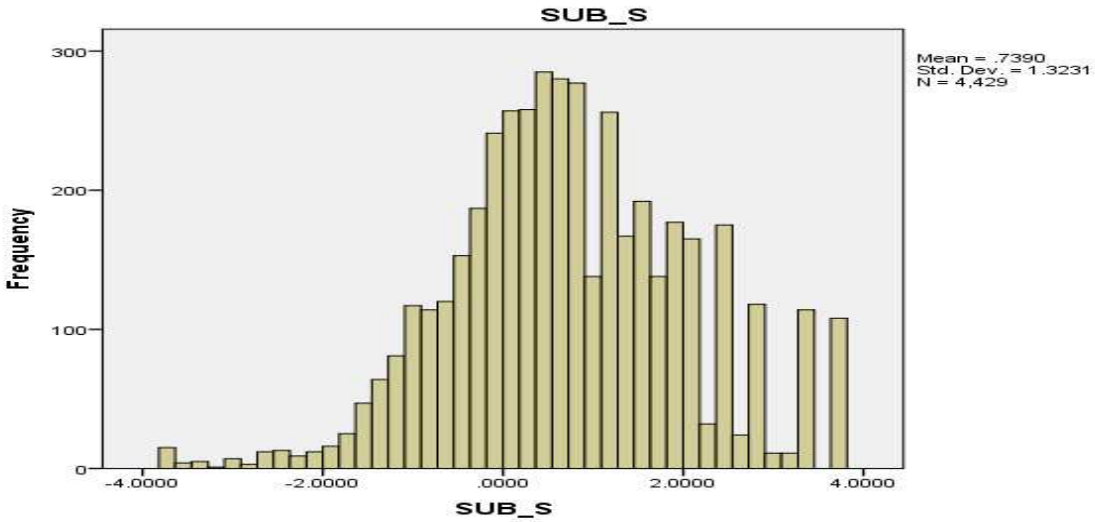

Figure 17. Speaking domain subscore distribution for grade 1 estimated with UIRT.

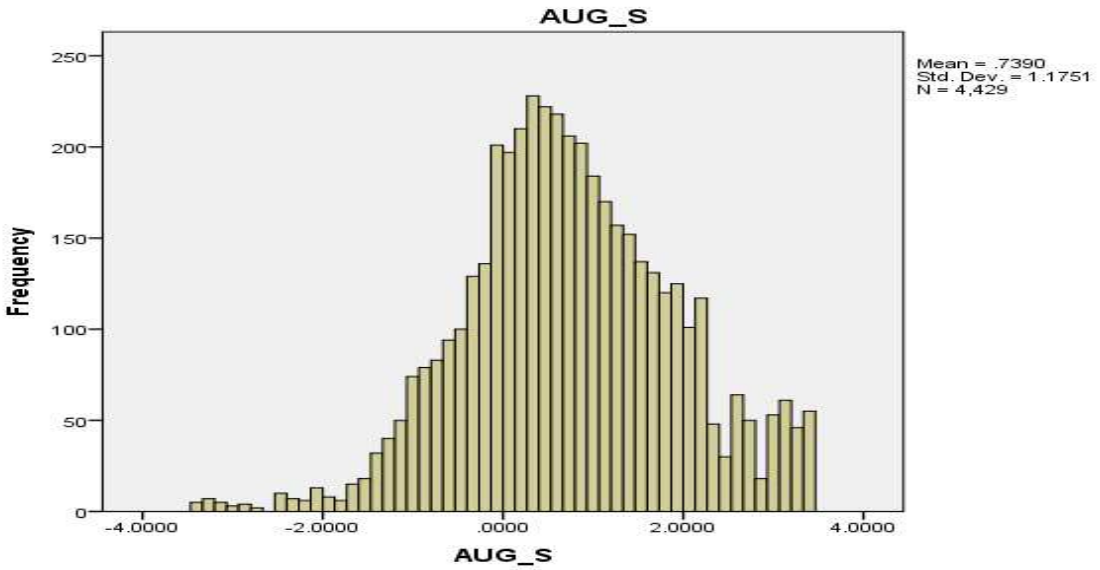

Figure 18. Speaking domain subscore distribution for grade 1 estimated with augmented IRT.

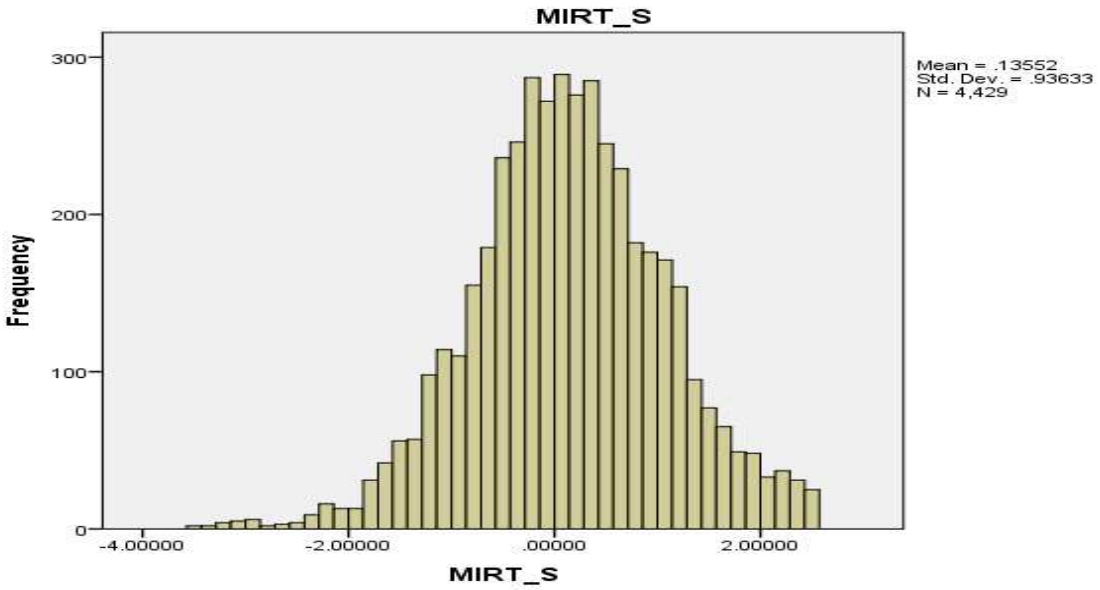

Figure 19. Speaking domain subscore distribution for grade 1 estimated with MIRT.

Subscore Correlations

Our first research question pertained to the correlations between subscores derived by four different methods:

1. How do the four methods of subscore reporting for four language domains compare in terms of subscore reliability and subscore correlations across five grade bands?

To investigate if the four methods were consistent with each other in assigning subscores, we performed subscore correlations between the four methods of subscore assignment for each grade band and domain. We found a high level of consistency in subscore assignment between estimation methods, ranging from .71 to .99. Generally, lower correlations were noted between MIRT and CTT (ranging from .75 to .99, with an average correlation of .93), CTT and augmented IRT (ranging from .82 to .99, with an average of .93), and between MIRT and UIRT (ranging from .71 to .99, with an average of .93). The highest correlations were noted between augmented IRT and MIRT (ranging from .73 to .99, with an average of .97). The example of grade 1 correlations between the methods is provided in table 10.

	CTT_L	IRT_L	AUG_L
CTT_L	1		
IRT_L	.96	1	
AUG_L	.88	.90	1
MIRT_L	.90	.90	.98
CTT_R	1		
IRT_R	.90	1	
AUG_R	.90	.99	1
MIRT_R	.93	.96	.98
CTT_W	1		
IRT_W	.98	1	

AUG_W	.89	.90	1
MIRT_W	.94	.94	.99
CTT_S	1		
IRT_S	.98	1	
AUG_S	.98	.99	1
MIRT_S	.98	.99	.99

Table 10. Grade 1 Subscore Correlations between the Scoring Methods.

Correlations between domains for CTT and UIRT methods were very similar across grade bands; correlations increased consistently when augmented IRT and MIRT methods were used. Correlations between domains ranged from .37 to .78 for CTT, from .37 to .75 for UIRT, from .55 to .98 for augmented IRT, and from .66 to .97 for MIRT.

Within grades, one consistent finding was a high correlation between reading and MC writing, ranging from .66 to .78 for CTT, from .64 to .75 for UIRT, from .90 to .98 for augmented IRT, and from .94 to .97 for MIRT. Prior to grade 2, the correlation between listening and reading was lower than the correlation between listening and MC writing in most cases. From grade 2 on, the correlation between listening and reading and listening and MC writing was fairly high, whereas the correlation between listening and speaking, reading and speaking, speaking and MC writing, speaking and writing rubric was lower. Between domains, correlations were almost always higher in grades 9-12 than in grade 0; however, the change in correlation was not always consistent between grades and between methods. For example, some correlations did increase fairly consistently (e.g., the correlation between listening and reading, reading and speaking, writing rubric and listening); others decrease first and then increase again (e.g. the

correlation between listening and writing, listening and speaking, speaking and writing, writing rubric and speaking), or did not change dramatically throughout the grade bands (e.g., the correlation between reading and writing, writing rubric and writing, writing rubric and reading). Table 11 illustrates correlations between domains within each method of subscore estimation for all grades.

Grade Band	Method	LR	LW	LS	RW	RS	SW	WR-L	WR-R	WR-W	WR-S
0	CTT	.49	.46	.49	.66	.40	.39	-	-	-	-
	UIRT	.48	.47	.50	.69	.39	.39	-	-	-	-
	A- IRT	.72	.80	.73	.95	.47	.55	-	-	-	-
	MIRT	.73	.79	.73	.94	.48	.55	-	-	-	-
1	CTT	.50	.55	.48	.70	.37	.42	-	-	-	-
	UIRT	.50	.54	.46	.70	.37	.42	-	-	-	-
	A- IRT	.79	.92	.72	.96	.47	.60	-	-	-	-
	MIRT	.81	.87	.70	.95	.50	.58	-	-	-	-
2	CTT	.55	.53	.43	.75	.38	.37	.42	.51	.51	.45
	UIRT	.56	.53	.41	.75	.38	.37	.42	.51	.51	.44
	A- IRT	.86	.85	.62	.98	.50	.51	.64	.67	.68	.56
	MIRT	.86	.85	.64	.96	.50	.51	.65	.65	.67	.57
3	CTT	.56	.52	.42	.77	.40	.39	.40	.54	.53	.44
	UIRT	.54	.50	.36	.73	.37	.35	.39	.54	.51	.42
	A- IRT	.84	.82	.57	.96	.49	.49	.61	.68	.69	.53
	MIRT	.88	.87	.62	.97	.51	.51	.66	.69	.72	.55
4-5	CTT	.58	.52	.42	.67	.41	.40	.41	.54	.50	.48
	UIRT	.56	.49	.36	.64	.38	.34	.39	.53	.47	.45
	A- IRT	.89	.88	.57	.96	.52	.52	.62	.69	.69	.57
	MIRT	.90	.89	.64	.96	.57	.59	.65	.71	.73	.64
6-8	CTT	.68	.59	.46	.72	.47	.43	.47	.54	.50	.45
	UIRT	.62	.53	.36	.65	.37	.33	.42	.49	.46	.40
	A- IRT	.91	.90	.54	.95	.48	.51	.62	.63	.68	.50
	MIRT	.94	.93	.66	.97	.63	.63	.68	.70	.71	.64
9-12	CTT	.67	.65	.54	.78	.56	.59	.50	.58	.56	.55
	UIRT	.61	.57	.42	.68	.44	.46	.45	.51	.49	.47
	A- IRT	.88	.87	.60	.90	.55	.60	.64	.64	.64	.57
	MIRT	.93	.93	.69	.97	.66	.70	.70	.71	.73	.71

Table 11. Correlations between Domains within Each Method of Subscore Estimation for All Grade Bands.

Subscore Reliability

Our first and second research questions pertained to the reliability and precision of the four methods of subscore reporting:

1. How do the four methods of subscore reporting for four language domains compare in terms of subscore reliability and subscore correlations across five grade bands?
2. How do the four methods of subscore reporting for four language domains compare in terms of subscore precision (standard error of measurement and conditional standard error of measurement) across five grade bands?

Overall, augmented IRT and MIRT had substantially higher reliability across all grades and all domains compared to the reliability of UIRT and CTT. We found that CTT reliability and precision was similar to UIRT reliability, with average reliability for both CTT and UIRT for listening at .66, reading - .85, MC writing - .72, speaking - .88, and writing rubric - .90. In addition, while the reliability of domains varied substantially between domains within a given grade for CTT and UIRT, it was more uniform (the reliability values for domains were closer together) in augmented IRT and MIRT. We found that augmented IRT was the most reliable method of subscore reporting, followed by MIRT, for all domains and grade bands. The average reliability of augmented IRT subscores was for listening at .85, for reading at .90, for MC writing at .89, for speaking at .89, and for writing rubric at .91. The average reliability of MIRT subscores was for listening, .76; for reading, .87; for MC writing, .82; for speaking, .84, and for writing rubric, .89. If one was to rank the reliability of domains across grades, fairly consistently writing rubric has the highest reliability, followed by speaking for all methods except for augmented IRT (in which case it is reading), then followed by reading, MC writing, and

listening. The least reliable domains received the highest increase in reliability when augmented IRT and MIRT were used. Tables 12-16 illustrate reliability of the four methods of subscore assignment for individual domains.

Standard Error of Measurement

Standard error of measurement (SE) was higher for CTT and UIRT than for augmented IRT and MIRT. For UIRT, the SE was lower than for CTT across all domains and grade bands. And, augmented IRT subscore average SE was consistently lower than the UIRT average SE. MIRT average SE was consistently lower than the UIRT average SE, and was also almost always lower than the augmented IRT average SE. The only two instances where the MIRT average SE was higher than the augmented IRT average SE was the listening and writing domains of grade 0. Tables 12-16 provide SE values of the four methods of subscore assignment for individual domains.

For all methods, the SE was generally decreasing from grade 0 to grade band 9-12 for most domains, although there were some grade bands that indicated an increase in SE. Across grade bands, for CTT, average SE was highest for the reading domain (1.82), followed by the listening domain (1.79), speaking (1.63), MC writing (1.38), and lowest for the writing rubric across all grade levels (1.09). For UIRT, the highest SE was in the listening domain (.94), followed by the MC writing domain (.75), reading (.54), speaking – .53; and writing rubric – .48. Average SE was fairly similar between UIRT, augmented IRT, and MIRT for reading, speaking, and writing rubric domains. For the writing and listening domains, UIRT average SE was higher than for other domains, and closer to the CTT SE. The average SE decrease was less consistent for augmented IRT (e.g., the augmented IRT SE for speaking was higher than that for reading,

but the reading reliability was lower than that for speaking). Also, while MIRT reliability was consistently lower than augmented IRT reliability, the MIRT average SE was consistently lower than that of augmented IRT. The average SE for augmented IRT for listening was .51; for reading and writing – .40; for speaking – .46; for writing rubric – .42. The average SE for MIRT for listening was .45; for reading – .33; for MC writing – .39; for speaking – .31; and for writing rubric, .28.

Domain Subscore Reliability

Listening Domain

Listening was consistently the least reliable domain across all grades and all methods, and consequently it received the most significant boost in reliability (a 21% increase in reliability from UIRT to augmented IRT on average across all grade bands). The increase in reliability from UIRT to MIRT was 13%. While CTT and UIRT reliability had more variation across grades and ranged from .58 to .72, augmented IRT and MIRT reliability was less varied and ranged from .83 to .86 and from .74 to .78, respectively (see table 12).

Method		Gr0	Gr1	Gr2	Gr3	Gr45	Gr68	Gr912	Average
CTT	reliability	.58	.62	.69	.67	.65	.70	.72	.66
	SE	2.14	1.85	1.87	1.67	1.64	1.62	1.74	1.79
UIRT	reliability	.58	.62	.69	.67	.65	.70	.72	.66
	avg SE	1.11	.99	.94	.94	.99	.83	.80	.94
A- IRT	reliability	.86	.85	.84	.83	.83	.86	.86	.85
	avg SE	.49	.48	.55	.54	.53	.48	.48	.51

MIRT	reliability	.77	.76	.78	.75	.74	.78	.78	.76
	avg SE	.49	.48	.48	.49	.46	.39	.34	.45

Table 12. Reliability and Standard Error (SE) for the Listening Domain for All Methods of Subscore Estimation and All Grade Bands. Average SE was calculated for IRT-based methods of subscore assignment.

MC Writing Domain

The initial CTT reliability of multiple choice (MC) writing was fairly high, so the average increase in reliability after augmentation was 19%; the increase in reliability when MIRT was used was 12%. While CTT and UIRT reliability had more variation across grades and ranged from .63 to .80, augmented IRT and MIRT reliability was less varied and ranged from .87 to .91 and from .79 to .86, respectively (see table 13).

Method		Gr0	Gr1	Gr2	Gr3	Gr45	Gr68	Gr912	Average
CTT	reliability	.63	.69	.78	.78	.67	.68	.80	.72
	SE	1.35	1.37	1.57	1.39	1.30	1.50	1.19	1.38
UIRT	reliability	.63	.69	.78	.78	.67	.68	.80	.72
	avg SE	.84	.82	.71	.60	.89	.80	.61	.75
A- IRT	reliability	.91	.90	.90	.90	.87	.88	.89	.89
	avg SE	.31	.38	.42	.42	.46	.42	.42	.40
MIRT	reliability	.82	.83	.86	.85	.79	.79	.81	.82
	avg SE	.43	.41	.38	.39	.41	.39	.30	.39

Table 13. Reliability and Standard Error (SE) for the Writing Domain for All Methods of Subscore Estimation and All Grade Bands. Average SE was calculated for IRT-based methods of subscore assignment.

Reading Domain

The initial CTT reliability of reading was fairly high, so the average increase in reliability after augmentation was 5%; the average increase in reliability when MIRT was used was 2%. Similar to other domains, while CTT and UIRT reliability had more variation across grades and ranged from .81 to .87, augmented IRT and MIRT reliability was less varied and ranged from .89 to .91 and from .85 to .91, respectively. However, even CTT and UIRT reliability had a fairly narrow range at this high level of reliability (see table 14).

Method		Gr0	Gr1	Gr2	Gr3	Gr45	Gr68	Gr912	Average
CTT	reliability	0.87	0.87	0.83	0.85	0.81	0.84	0.86	0.85
	SE	1.88	1.53	2.06	1.86	1.86	1.79	1.78	1.82
UIRT	reliability	0.87	0.87	0.83	0.85	0.81	0.84	0.86	0.85
	avg SE	0.44	0.53	0.61	0.53	0.62	0.51	0.53	0.54
A- IRT	reliability	0.90	0.90	0.90	0.91	0.89	0.90	0.91	0.90
	avg SE	0.35	0.42	0.41	0.39	0.43	0.38	0.41	0.40
MIRT	reliability	0.91	0.87	0.88	0.88	0.85	0.85	0.85	0.87
	avg SE	0.30	0.36	0.36	0.35	0.35	0.33	0.28	0.33

Table 14. Reliability and Standard Error (SE) for the Reading Domain for All Methods of Subscore Estimation and All Grade Bands. Average SE was calculated for IRT-based methods of subscore assignment.

Speaking Domain

The initial CTT reliability of reading was high, so the average increase in reliability after augmentation was only 1%; and average reliability decreased when MIRT was used. Similar to other domains, while CTT and UIRT reliability had more variation across grades and ranged

from .86 to .91, augmented IRT and MIRT reliability was less varied and ranges from .88 to .91 and from .74 to .92, respectively. All methods' reliability had a fairly narrow range at this high level of reliability (see table 15).

Method		Gr0	Gr1	Gr2	Gr3	Gr45	Gr68	Gr912	Average
CTT	reliability	.91	.89	.87	.86	.87	.87	.91	.88
	SE	2.01	1.87	1.74	1.62	1.50	1.34	1.32	1.63
UIRT	reliability	.91	.89	.87	.86	.87	.87	.91	.88
	avg SE	.46	.47	.57	.62	.60	.57	.46	.53
A- IRT	reliability	.91	.89	.89	.88	.88	.88	.91	.89
	avg SE	.41	.41	.48	.51	.51	.48	.41	.46
MIRT	reliability	.92	.90	.89	.88	.80	.77	.74	.84
	avg SE	.28	.31	.33	.33	.33	.32	.27	.31

Table 15. Reliability and Standard Error (SE) for the Speaking Domain for All Methods of Subscore Estimation and all Grade Bands. Average SE was calculated for IRT-based methods of subscore assignment.

Writing Rubric

Writing rubric was consistently the domain with the highest reliability across all grades. The average increase in reliability after augmentation was only 1%; and average reliability decreased when MIRT was used. Similar to other domains, while CTT and UIRT reliability had more variation across grades and ranged from .89 to .91, augmented IRT and MIRT reliability was less varied and ranged from .90 to .92 and from .85 to .90, respectively. All methods' reliability had a fairly narrow range at this high level of reliability. Similarly to speaking, MIRT reliability in the writing rubric domain decreased at grade band level 4-5 (see table 16).

Method		Gr2	Gr3	Gr45	Gr68	Gr912	Average
CTT	reliability	.89	.89	.90	.89	.91	.90
	SE	1.16	1.13	1.09	1.06	1.00	1.09
UIRT	reliability	.89	.89	.90	.89	.91	.90
	avg SE	.51	.48	.50	.48	.45	.48
A- IRT	reliability	.90	.90	.91	.90	.92	.91
	avg SE	.44	.41	.44	.41	.39	.42
MIRT	reliability	.90	.90	.89	.88	.85	.89
	avg SE	.31	.31	.27	.27	.24	.28

Table 16. Reliability and Standard Error (SE) for the Writing Rubric Domain for All Methods of Subscore Estimation and all Grade Bands. Average SE was calculated for IRT-based methods of subscore assignment.

Relationship between Correlation and Reliability in Augmented IRT Subscores

Based on the correlations between the domains and the reliability of domains, in calculating augmented subscores, the B matrix of weights was constructed; these weights indicated the magnitude of impact of other domains on the domain of interest. This relationship between correlation and reliability in subscore augmentation can be illustrated as follows: listening had the highest correlation with speaking; speaking was also the most reliable domain. In addition, because speaking was substantially more reliable than listening, speaking had the weight of the highest magnitude in the B matrix for grade 0. For grade 1, writing had the highest correlation with listening. While speaking had the highest reliability, its correlation with listening was lower than that of writing. In addition, the reliability of listening itself increased in grade 1

as compared to grade 0. Therefore, listening itself had the biggest impact on the listening subscore, followed by writing. In grade bands 2, 3, 4-5, 6-8, and 9-12 the reliability of listening continued to increase, so listening carried the biggest weight in the augmented listening subscore. Listening correlated most highly with reading; while speaking had the highest reliability, reading carried the next highest weight in the impact on the augmented listening score.

Reading was a highly reliable domain; it consistently correlated most highly with writing, another domain with high reliability. Therefore, after itself, the highest impact on the augmented reading scores came from the writing domain. However, in grade band 6-8 the reliability of reading was higher than all other domains except for writing rubric. Since the correlation of reading with writing rubric was fairly high, writing rubric had a higher impact on augmented reading than the writing domain in band 6-8 (B=0.11).

The reliability of MC writing scores derived with the UIRT method varied from one grade band to another. However, writing correlated highly with a very reliable domain, reading. In grades 0 and 1, the reliability of writing was substantively lower than that of reading, therefore reading had the biggest impact, even bigger than the writing score itself, on the augmented writing score for these grades. In grade band 4-5, writing and reading scores had an equal impact on the augmented writing score. In other grade bands, since the reliability of writing became higher, writing itself had the biggest impact on the augmented writing score, followed closely by the reading score impact.

Speaking was a domain with high reliability; it also did not have very high correlations with any other domain. Therefore, throughout all grade bands, speaking itself had the highest impact on the augmented speaking score. Other domains shared a small amount of impact on the

speaking domain evenly. One exception was the high impact of writing rubric score on the augmented speaking domain in grade band 6-8 (B=.12). While not very high, the correlation of writing rubric and speaking scores was the highest that speaking had with any other domain for this grade band, and the reliability of writing rubric was very high.

Writing rubric was the domain with the highest reliability, and therefore UIRT derived writing rubric score had the highest impact on the writing rubric augmented score. The impact of other domains was fairly small, with the exception of reading impact on the writing rubric in grade band 6-8 (B=.18). Reading was the domain with which writing rubric UIRT score correlates most highly, and it was a highly reliable domain. Tables 17-23 contain the B weights used to calculate augmented IRT scores for grade bands and domains.

	L	R	W	S
L	.20	.22	.10	**.28**
R	.02	**.63**	.08	.02
W	.05	**.54**	.14	.07
S	.02	.02	.01	**.80**

Table 17. Grade 0 B Weights Used to Calculate Augmented IRT Scores from UIRT Scores. The largest weight in each row is emphasized in boldface.

	L	R	W	S
L	**.24**	.13	.15	.20
R	.02	**.64**	.10	.01
W	.08	**.37**	.21	.09
S	.02	.01	.02	**.76**

Table 18. Grade 1 B Weights Used to Calculate Augmented IRT Scores from UIRT Scores. The largest weight in each row is emphasized in boldface.

	L	R	W	S	WRUB
L	**.34**	.16	.09	.10	.05
R	.05	**.46**	.19	.01	.05
W	.04	.29	**.35**	.01	.07
S	.02	.01	.01	**.72**	.05
WRUB	.01	.03	.03	.04	**.74**

Table 19. Grade 2 B Weights Used to Calculate Augmented IRT Scores from UIRT Scores. The largest weight in each row is emphasized in boldface.

	L	R	W	S	WRUB
L	**.33**	.18	.09	.08	.03
R	.04	**.53**	.14	.01	.06
W	.03	.28	**.38**	.01	.07
S	.02	.02	.01	**.70**	.05
WRUB	.00	.04	.03	.03	**.74**

Table 20. Grade 3 B Weights Used to Calculate Augmented IRT Scores from UIRT Scores. The largest weight in each row is emphasized in boldface.

	L	R	W	S	WRUB
L	**.29**	.21	.07	.07	.03
R	.05	**.48**	.09	.02	.08
W	.06	**.27**	.27	.02	.09
S	.01	.01	.01	**.72**	.05
WRUB	.00	.04	.02	.03	**.76**

Table 21. Grade Band 4-5 B Weights Used to Calculate Augmented IRT Scores from UIRT Scores. The largest weight in each row is emphasized in boldface.

	L	R	W	S	WRUB
L	**.42**	.28	.06	.07	.04
R	.04	**.69**	.09	.01	.11
W	.05	**.41**	.31	.02	12
S	.01	.07	-.01	**.76**	.12
WRUB	-.04	.18	.00	.06	**.85**

Table 22. Grade Band 6-8 B Weights Used to Calculate Augmented IRT Scores from UIRT Scores. The largest weight in each row is emphasized in boldface.

	L	R	W	S	WRUB
L	**.36**	.17	.11	.06	.05
R	.05	**.59**	.11	.02	.05
W	.05	.16	**.46**	.05	.04
S	.01	.01	.02	**.79**	.03
WRUB	.01	.03	.02	.03	**.78**

Table 23. Grade Band 9-12 B Weights Used to Calculate Augmented IRT Scores from UIRT Scores. The largest weight in each row is emphasized in boldface.

Correlations between Domains in the MIRT Framework

MIRT estimates item parameters (slope and intercept) by direct estimation of the variance– covariance matrix (covariances are listed in table 24) of latent traits, and the loading of each item on the specified domain.

Grade Band	LR	LW	LS	RW	RS	SW	WR-L	WR-R	WR-W	WR-S
0	.61	.65	.62	.85	.42	.47	-	-	-	-
1	.70	.75	.61	.87	.46	.52	-	-	-	-
2	.73	.71	.55	.88	.44	.44	.54	.57	.59	.52
3	.76	.74	.55	.89	.48	.48	.57	.64	.65	.53
4-5	.82	.79	.63	.89	.60	.60	.64	.71	.71	.68
6-8	.88	.85	.69	.91	.68	.67	.71	.75	.73	.72
9-12	.90	.90	.79	.94	.79	.80	.81	.82	.83	.85

Table 24. Domain Covariances by Grade Band Estimated by MIRT.

A confirmatory MIRT model used in this study can be described as a full information confirmatory factor analysis. First, the specification is made regarding which items assess a specific domain. MIRT factor loadings, similar to factor analysis loadings, indicate the weight of specific items in determining the proficiency in the specified domain. The slope and intercept parameters (see examples in tables 2-3 of Appendix 1) were calculated from the factor loadings and the estimated covariance matrix between dimensions, and then using these values, the student θ values were calculated. The magnitude of covariances was generally close to the magnitude of correlations of subscores estimated with other methods.

Subscore Variability and Precision at Different Proficiency Levels

Our third research question was as follows:

3. How does the precision of the four methods of subscore reporting in the four language domains compare in the estimation of ability of examinees within five grade bands at different proficiency levels?

The percentage of students at proficiency levels 1 (beginning), 2 (intermediate), and 3 (advanced) decreased from grade 0 to grade band 9-12. While there were still some students at proficiency level 1 in grade band 9-12, the majority of students were more proficient. Conversely, the percentage of students at proficiency level 4 (advanced) increased from grade 0 to grade band 9-12.

To be able to evaluate subscore variability and precision, we calculated the average SD and SE for CTT, UIRT, and MIRT for the scores at given proficiency levels. The augmented IRT SE is averaged across all scores, and therefore individual score SE is not available. We then compared the SD and average SE at specific proficiency levels to determine if some proficiency

levels had a substantial difference in those values. We used R package *'mirt'* (Chalmers, 2012) to estimate SE for UIRT subscores.

Overall, across domains and grades the standard deviation (SD) for all methods of subscore assignment becomes lower if there are more students at a given proficiency level, and if the range of possible points for answering the assessment questions is narrow. The range of scores was the highest at proficiency level 1; at the same time, the percentage of students at level 1 was the lowest. Conversely, since the range of scores continued to decrease at proficiency levels 2-4, but the number of students at those levels was higher than at level 1, the SD values were lower for those proficiency levels. IRT-estimated SD values were consistently close together, with MIRT having the lowest SD of the three IRT-derived subscores, and UIRT – the highest SD values. CTT-derived subscores consistently have the highest SD of all four methods of subscore estimation (see grade 1 example in table 25).

SE in CTT was the highest at level 1, and at times at level 2, and decreased in levels 3 and 4 across all grade bands and domains. IRT SE either stayed approximately the same (writing rubric), increased from level 1 to level 4 (speaking, listening, writing), or decreased from level 1 to levels 2 and 3 and then slightly increased again at level 4 (reading). Generally, proficiency level 4 had the highest IRT SE and lowest CTT SE across domains and grade bands. IRT-estimated SE values were consistently close together, with MIRT having the lowest SE of the three IRT-derived subscores, and UIRT – the highest SE values. CTT-derived subscores consistently have the highest SE of all four methods of subscore estimation (see grade 1 example in table 25).

We encountered some inconsistencies when calculating CTT subscore correlations at specific proficiency levels, with some correlations becoming very small or negative after the

scores were grouped by total proficiency levels. This situation best fits the pattern of Berkson's paradox (Elwert & Winship, 2014; Westreich, 2012; Pearl, 2000), in which the assessment of the relationship between the domains becomes biased due to the fact that the total proficiency level was conditioned on proficiency in the individual domains.

Due to CTT subscore range restriction at specific proficiency levels, and consequently due to very limited variance of student (total) scores, as opposed to a more significant variance of item scores, very low or even negative values for CTT reliability (Cronbach's alpha) of subscores at specific proficiency levels resulted. Those values were substantially different from the reliability of domain subscores as a whole calculated for students at all levels of proficiency. Therefore those reliability values for specific proficiency levels were not compared with UIRT and MIRT reliability for specific proficiency levels In addition, since the reliability of augmented subscores is calculated as the average reliability for a specific domain, the reliability of augmented subscores at specific proficiency levels was not calculated.

Gr1	method	Prof 1	Prof 2	Prof 3	Prof 4
Listening	CTT reliability (avg)	.62	.62	.62	.62
	CTT SD	1.65	1.32	.50	.48
	CTT SE	2.37	2.18	1.75	1.26
	UIRT reliability	.50	.47	.33	.25
	UIRT SD	.48	.52	.45	.40
	UIRT SE	.47	.55	.64	.70
	A- IRT reliability (avg)	.85	.85	.85	.85
	A-IRT SD	.50	.48	.46	.45
	MIRT reliability	.55	.51	.41	.45

	MIRT SD	.46	.47	.42	.50
	MIRT SE	.41	.46	.51	.56
Reading	CTT reliability (avg)	.87	.87	.87	.87
	CTT SD	2.75	1.43	.81	.50
	CTT SE	2.42	2.22	1.60	.43
	UIRT reliability	.76	.48	.47	.49
	UIRT SD	.46	.25	.35	.59
	UIRT SE	.25	.26	.37	.59
	A- IRT reliability (avg)	.90	.90	.90	.90
	A-IRT SD	.44	.27	.34	.53
	MIRT reliability	.75	.49	.42	.44
	MIRT SD	.41	.24	.28	.43
	MIRT SE	.24	.24	.33	.48
Speaking	CTT reliability (avg)	.89	.89	.89	.89
	CTT SD	3.55	1.88	1.66	1.38
	CTT SE	2.47	2.78	2.41	1.39
	UIRT reliability	.88	.60	.68	.70
	UIRT SD	.71	.36	.46	.58
	UIRT SE	.27	.29	.32	.38
	A- IRT reliability (avg)	.89	.89	.89	.89
	A-IRT SD	.63	.33	.41	.52
	MIRT reliability	.79	.51	.52	.59
	MIRT SD	.51	.29	.32	.43

	MIRT SE	.26	.28	.31	.36
Writing	CTT reliability (avg)	.69	.69	.69	.69
	CTT SD	1.36	.81	.49	.00
	CTT SE	1.90	1.73	1.21	.00
	UIRT reliability	.54	.39	.34	.01
	UIRT SD	.55	.42	.42	.00
	UIRT SE	.51	.52	.58	.66
	A- IRT reliability (avg)	.90	.90	.90	.90
	A-IRT SD	.57	.53	.46	.31
	MIRT reliability	.66	.50	.40	.27
	MIRT SD	.53	.40	.38	.34
	MIRT SE	.38	.40	.47	.55

Table 25. Grade 1 Subscore Reliability by Proficiency Level for Four Subscore Estimation Methods. Average reliability is listed for CTT and augmented IRT subscores. Values of 0.00 are due to rounding.

While cut scores are available for writing as a domain, they are not available separately for writing multiple choice and writing rubric items, so we could only approximate the proficiency levels for these two parts of the writing domain. Figures 21-23 illustrate CSEM for the four proficiency levels for grade 1 listening domain for four methods of subscore assignment. In the CTT figure, the cut scores for proficiency levels are depicted; for IRT-derived subscores, the mean ability for proficiency levels are depicted.

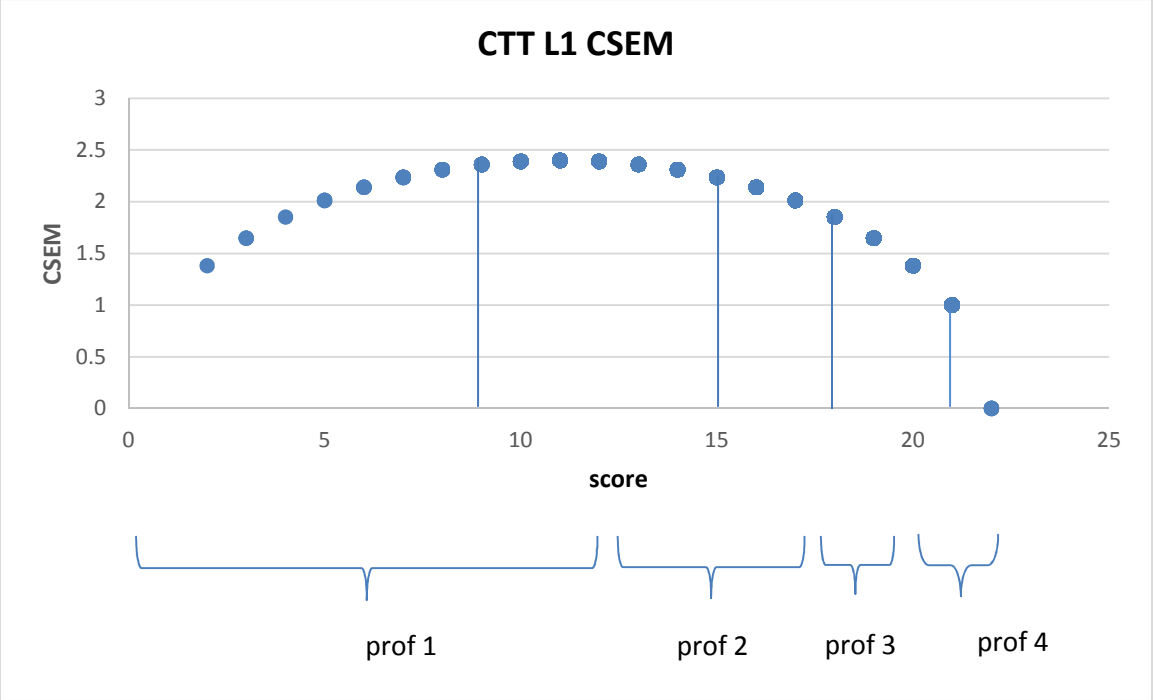

Figure 20. Grade 1 listening domain CTT CSEM by proficiency level.

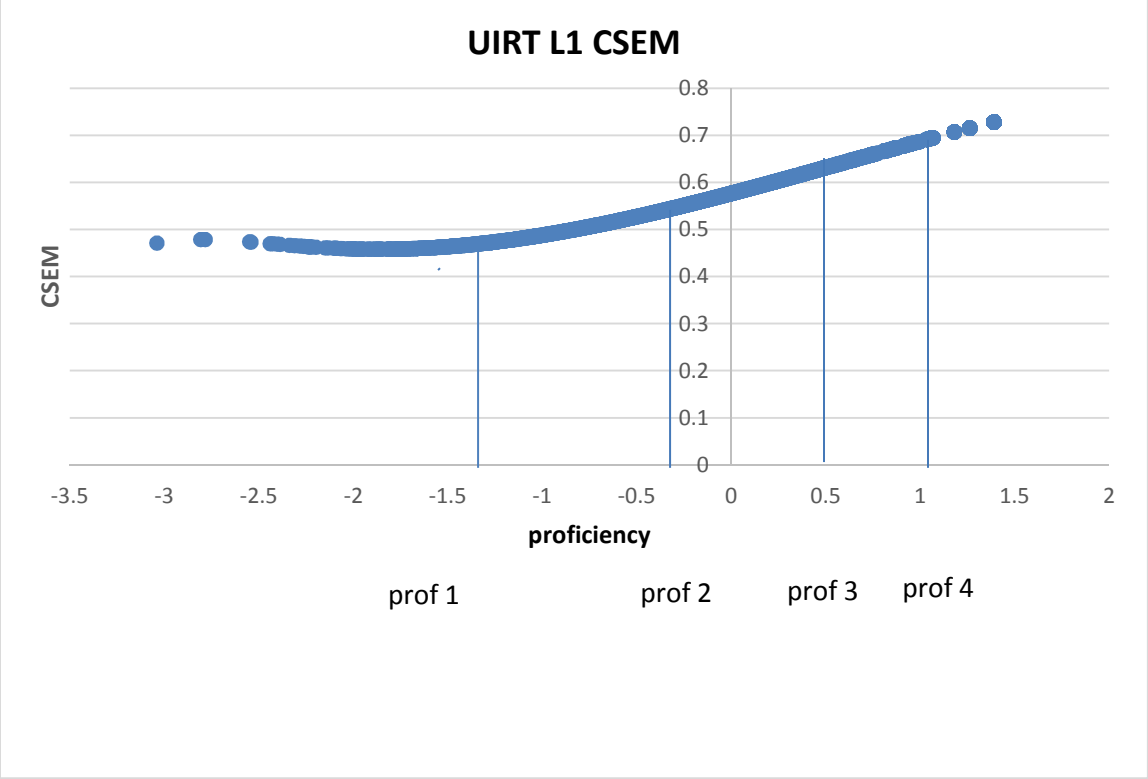

Figure 21. Grade 1 listening domain UIRT CSEM by proficiency level.

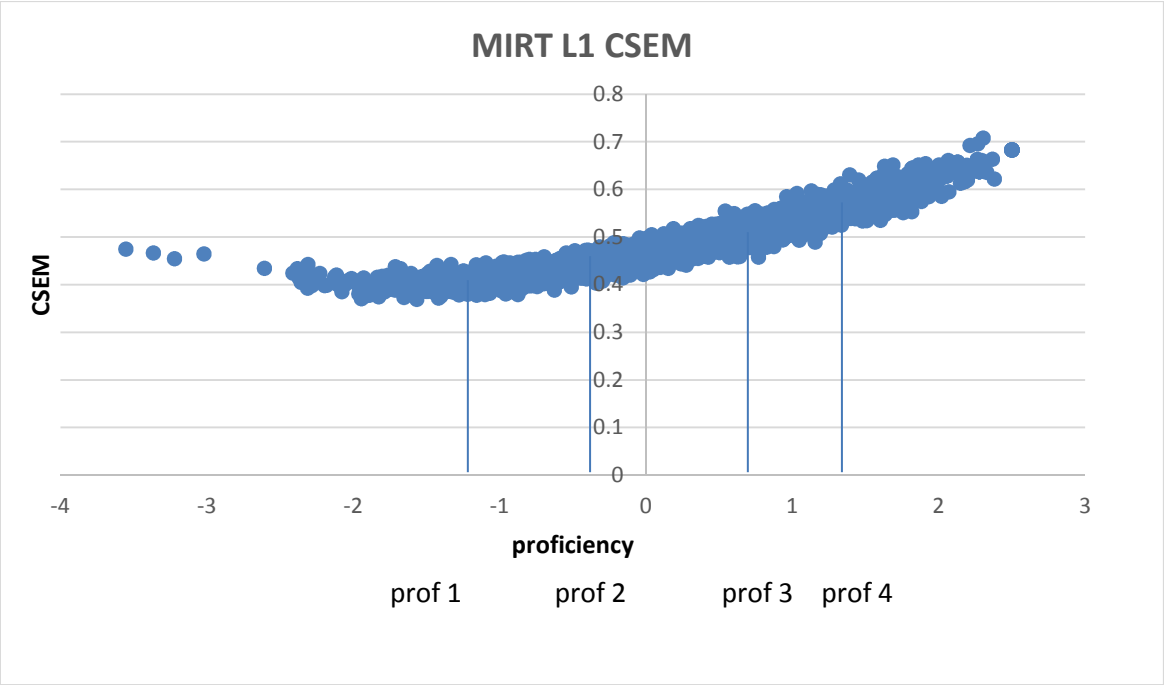

Figure 22. Grade 1 listening domain MIRT CSEM by proficiency level.

CHAPTER 5

DISCUSSION

In this chapter, we discuss the domain subscore descriptives, correlations between domains, domain reliability of subscores obtained with different methods, and the impact of different methods of subscore assignment on subscore interpretation. This chapter interprets the findings and presents implications for educational policy and practice, suggestions for future research, and limitations of the study.

Descriptives

The four domains of the state English language proficiency test (listening, reading, speaking, and writing) were scored with four methods (classical test theory (CTT), or number correct; unidimensional item response theory (UIRT); augmented IRT; and multidimensional IRT (MIRT). First, the descriptives for all four methods, four domains, and grade bands were obtained. The means within the same domain were increasing across all domains from K to grade band 9-12. The mean for UIRT subscores was the same as the mean for the augmented IRT subscores, as augmented IRT scores are linear transformations of UIRT scores. The means of IRT-derived scores for all domains and grade bands were close to 0, which is the mean of normally distributed ability. On average the CTT mean for listening was 16.84; for reading – 17.80, for multiple choice (MC) writing – 11.02; for speaking – 24.50, and for writing rubric – 14.04. On average the UIRT and augmented IRT mean for listening was .70, for reading - .74; for writing - .78; for speaking – 1.18; for writing rubric - .88. For MIRT, the average mean for listening was .20; for reading - .21; for MC writing - .21; for speaking - .18; and for writing rubric - .27. Since the normality of score distribution was retained in MIRT estimation, the

means of these subscores are close to 0. The mean of subscores did not change between UIRT and augmented IRT; however, the means are different between domains. For MIRT-derived subscores, the means are closer together across domains.

The variability of subscores was generally shown to be reduced from CTT to UIRT to augmented IRT to MIRT. Part of the reason for smaller variance of IRT-based scores was the estimation method used (EAP) that underestimates variability and makes individuals' scores more like the group mean. Generally, the more points were possible in a domain, the higher the SD was for that domain. For CTT subscores, the average SD for listening was 3.01; for reading – 4.67; for writing – 2.66, and for writing rubric – 3.37. For UIRT subscores, the SD was lower than for CTT across all domains and grade bands. The average SD for UIRT subscores for listening was 1.32; for reading – 1.27; for multiple choice (MC) writing – 1.21; for speaking – 1.46; for writing rubric – 1.42. Augmented IRT subscore SD was consistently lower than the UIRT SD. Augmented IRT subscore SD was consistently lower than the UIRT SD. The average SD for augmented IRT subscores for listening was 0.94; for reading – 1.10; for MC writing – .96; for speaking – 1.30; and for writing rubric – 1.28. MIRT subscore SD was consistently lower than the UIRT subscore SD, and was also almost always lower than the augmented IRT subscore SD. The average SD for MIRT subscores for listening was .81; for reading – .86; for MC writing – .84; for speaking – .78; and for writing rubric, .80.

Augmented IRT subscores of any given individual were closer to the mean of the group subscores. On the other hand, MIRT subscores looked flatter in general, since the subscores of an individual for a given domain were more like this individual' subscores for other domains. For the five participants randomly selected for visual presentation purposes, the highest variability between different domains was for the subscores obtained with the CTT and UIRT methods; this

finding was consistent across all grade bands. The boxplots constructed for the domains and grade levels supported this conclusion as well. The overall reduction of variability in augmented scores was supported by a number of studies (e.g., Wainer et al., 2001; Skorupski, 2008). The reduction in variability from UIRT to MIRT was also supported by studies examining different methods of subscore assignment (e.g., Dwyer et al., 2006; de la Torre, Song, & Hong, 2011; de la Torre & Song, 2009). While MIRT-based subscore distribution tended to be more normal, CTT-based subscore distribution tended to be negatively skewed for most grade bands, indicating that the majority of students were able to answer most of the items correctly.

Correlations

Correlations between Subscores Estimated by Different Methods

We found that there was a high level of consistency in subscore assignment between estimation methods, ranging from .71 to .99. Generally, lower correlations were noted between MIRT and CTT (ranging from .75 to .99, with an average correlation of .93), CTT and augmented IRT (ranging from .82 to .99, with an average of .93), and between MIRT and UIRT (ranging from .71 to .99, with an average of .93). The highest correlations were noted between augmented IRT and MIRT (ranging from .73 to .99, with an average of .97). This finding was expected, since augmented IRT derived subscores have more characteristics in common with MIRT (such as the IRT-based framework, and the utilization of information from other subscores) than with CTT-derived subscores.

Correlations between Domains across Grade Bands

Within pairs of domains, correlations were almost always higher in grades 9-12 than in grade 0; however, the increase in the magnitude of the correlation was not always consistent between grades and between methods. For example, correlations between some domain pairs did increase fairly consistently (e.g. the correlation between listening and reading increased from .49 in grade 0 to .67 in grade band 9-12 for CTT, from .72 in grade 0 to .88 in grade band 9-12 for augmented IRT in grade band 9-12; the correlation between reading and speaking increased from .40 in grade 0 to .56 in grade band 9-12 in CTT, and from .47 in grade 0 to .55 in grade band 9-12 for augmented IRT); others decreased first and then increased again (e.g., the correlation between listening and speaking decreased from .49 in grade 0 to .42 in grade band 4-5, and then increased to .54 in grade band 9-12 in CTT), or did not change dramatically throughout the grade bands (e.g., the correlation between reading and MC writing stayed consistently high, ranging from .66 to .78 for CTT, from .64 to .75 for UIRT, from .90 to .98 for augmented IRT, and from .94 to .97 for MIRT). Other studies confirm that there is little consistency in correlations between different language domains in both cross-sectional and longitudinal studies (Farnia & Geva, 2013; Kieffer, 2011). It appears, however, that one correlation that was most consistently high across grades was the correlation between reading and MC writing, with an average of .72 for CTT, .69 for UIRT, .95 for augmented IRT, and .96 for MIRT. This finding is supported by previous studies exploring relationships between language domains, although the magnitude of this relationship and the direction of causality (i.e., whether changes in reading skills cause changes in writing skills, or changes in writing skills cause changes in reading skills) is not consistent from study to study (Davis & Bryant, 2006; Forrester, 2013; Shanahan & Lomax, 1986, 1988; Farnia & Geva, 2013; Manis, Lindsey, & Bailey, 2004; Hammill, 2004; Scarborough, 1998).

To a degree, it was somewhat surprising that the correlation between MC writing and writing rubric was only averaging .52 for CTT subscores, and .49 for UIRT. However, the average correlation between these two subscores increased for augmented IRT to an average of .68, and for MIRT – to an average of .71. It is also likely that the magnitude of this correlation had to do with different methods with which these domains were assessed. As mentioned previously, different tasks were used to assess MC writing and the writing rubric. The MC section included the identification of grammatically correct use of parts of speech; identification of synonyms/antonyms and correct use of punctuation, grammar, and syntax in sentences. The writing rubric, on the other hand, involved writing an essay based on either a picture or a written prompt. For this task, students were evaluated based on the expressiveness of vocabulary; correct sentence structure; grammar; punctuation, spelling, and capitalization; and the general organization of the essay.

In addition, since correlations tend to be sample-specific, part of the reason for the low correlations could be due to the characteristics of the specific student sample. E.g., in this specific sample, the correlations between the writing multiple choice items and writing rubric were lower than may be expected possibly due to the fact that they were significantly more or less skilled in the task of writing several coherent paragraphs than in answering multiple choice questions; or they were significantly less familiar with the multiple choice task than with the task of writing a coherent passage; in another sample, the interaction between the students' ability and the task may be different.

Correlations between Domains across Estimation Methods

Correlations between domains for CTT and UIRT methods were very similar; correlations increased consistently when augmented IRT and MIRT methods were used. This

finding was expected; once augmentation procedures are implemented, correlation between subscores increases (Skorupski, 2008; Wainer et al., 2001; Stone et al., 2010; de la Torre, Song, & Hong, 2011). Correlations between domains ranged from .37 to .78 with an average of .52 for CTT, from .33 to .75 with an average of .48 for UIRT, from .47 to .98 with an average of .69 for augmented IRT, and from .48 to .97 with an average of .73 for MIRT.

Reliability of the Four Methods of Subscore Assignment

CTT vs. UIRT

We found that CTT reliability and precision was similar to UIRT reliability, with average reliability for both CTT and UIRT for listening at .66, reading - .85, MC writing - .72, speaking - .88, and writing rubric - .90. Some previously done studies came to the same conclusion regarding the similarity of CTT and UIRT reliability (e.g.. Dwyer et al., 2006; Haberman & Sinharay, 2010l Luecht, 2003). In a number of other studies, however, UIRT reliability was found to be higher than CTT reliability (Yao & Boughton, 2007; Shin, 2007; Ferrando and Chico, 2007; Xu & Stone, 2012). It is understandable that since these two methods, while different in their key assumptions, are still similar in the sense that they treat each domain as a separate test and do not use any additional information about student ability, internal or external to the test, the reliability of these two approaches tended to be similar. In addition, when the number of items in a domain is large, as was the case with the present assessment of language proficiency, these two methods often tend to have similar reliability.

UIRT vs. Augmented IRT

The reliability of augmented IRT was consistently found to be higher than that of non-augmented IRT. While the average reliability of UIRT was for listening at .66, reading - .85, MC writing - .72, speaking - .88, and writing rubric - .90, the average reliability of augmented IRT subscores was for listening at .85, for reading at .90, for MC writing at .89, for speaking at .89, and for writing rubric at .91. This finding regarding the increase in reliability from non-augmented to augmented subscores was supported by a number of previous studies (Wang, Chen & Cheng, 2004; Yao & Boughton, 2007; Haberman & Sinharay, 2010; De la Torre & Song, 2009; Wainer et al., 2001; Skorupski, 2008).

MIRT vs. CTT/ UIRT

MIRT was consistently found to be more reliable than the non-augmented methods of subscore estimation (CTT and UIRT). This finding is supported by other research investigating the reliability of these subscore estimation methods (Yao & Boughton, 2007; de la Torre & Patz, 2005; de la Torre, Song, & Hong, 2011). In assessments where correlations between domains are negligible, using MIRT vs. non-augmented methods would not make much difference in reliability. However, in assessments where correlations between domains were larger, MIRT was consistently found to have higher reliability.

MIRT vs. Augmented IRT

We found that while MIRT reliability and augmented IRT reliability were very close for most grade bands and domains, in the majority of cases, MIRT reliability was lower than augmented IRT reliability, despite the fact that SEM was consistently lower for MIRT-estimated subscores in all but two instances (MC writing domain for grade 0 and grade 1). The average

reliability of augmented IRT subscores was for listening at .85, for reading at .90, for MC writing at .89, for speaking at .89, and for writing rubric at .91. The average reliability of MIRT subscores was for listening, .76; for reading, .87; for MC writing, .82; for speaking, .84, and for writing rubric, .89. A number of studies have concluded that MIRT and augmented unidimensional IRT can have very similar reliability (e.g., Gessaroli, 2004; Luecht, 2003; Haberman & Sinharay, 2010), or that MIRT reliability may be higher than augmented IRT reliability (De la Torre & Patz 2005; Dwyer et al., 2006; De la Torre, Song, & Hong, 2011). Due to increased computational and interpretation demands of MIRT, this finding of similarity in reliability, or rather small gains in reliability, may serve as a ground to continue using computationally simpler methods, such as UIRT or augmented IRT, instead of MIRT.

The decrease of MIRT reliability at grade bands 4-5, 6-8, and 9-12 for the writing rubric (from .90 to .85) and speaking scores (from .88 to .74) was not consistent with the reliability changes for other three methods of subscore reporting; CTT, UIRT, and augmented IRT reliability increased in all domains, including speaking and writing rubric, by grade band 9-12. It appears that a possible explanation of this decrease is the decrease in the variability of MIRT-derived subscores. When reliability is measured as the ratio of variability to the sum of variability and squared standard error, the higher the ratio of squared standard error to variability, the lower the reliability is. Despite the fact that both standard error and variability are consistently lower in MIRT-derived subscores, and the correlations between all MIRT-derived subscores either remained similar or increased from grade 0 to grade band 9-12, the ratio of squared standard error to variability was higher for MIRT-derived subscores than for UIRT-derived subscores, for example. Also, while MIRT reliability was slightly lower than augmented IRT reliability, the MIRT average SE was also consistently lower than that of augmented IRT. Both the SE and the reliability for augmented IRT and MIRT were calculated with different

methods, which were conceptually related, but different in the formulas they used to measure SE and reliability. In addition, the formula used to calculate SE for augmented IRT was conceptually based on the CTT framework, whereas the SE calculated for MIRT was the average of SE values calculated based on the IRT framework. The reliability calculations for both methods, while conceptually related and based on the CTT framework, were also formulaically different. Augmented IRT reliability was calculated as the ratio of the unconditional true score variance to unconditional estimated true score variance (Wainer et al., 2001). MIRT reliability was calculated as the ratio of true score variance to the sum of true score variance and the average value of the variance of error measurement associated with those scores.

Because real data were used, it may not have been possible to say which subscore reporting method was most accurate. However, based on the calculations of reliability we were using for this study, we found that augmented IRT had the highest reliability of all four methods of subscore reporting, followed by MIRT, for all domains and grade bands, while the reliability of CTT and UIRT was the lowest.

Standard Error of Measurement

Standard error of measurement was consistently higher for CTT and UIRT than for augmented IRT and MIRT. For CTT, the average SE for listening was 1.79; for reading – 1.82; for MC writing – 1.38; for speaking – 1.63; and for writing rubric – 1.09. For UIRT, the SE was lower than for CTT across all domains and grade bands. The average SE for UIRT for listening was .94; for reading – .54; for MC writing – .75; for speaking – .53; and for writing rubric – .48. And, augmented IRT subscore average SE was consistently lower than the UIRT average SE. The average SE for augmented IRT for listening was .51; for reading and MC writing – .40; for

speaking – .46; for writing rubric – .42. MIRT average SE was consistently lower than the UIRT average SE, and was also almost always lower than the augmented IRT average SE. The only two instances where the MIRT average SE was higher than the augmented IRT average SE was the listening and MC writing domains of grade 0. The average SE for MIRT for listening was .45; for reading – .33; for MC writing – .39; for speaking – .31; and for writing rubric, .28.

For all methods, the SE was generally decreasing from grade 0 to grade band 9-12 for most domains, although there were some grade bands that had an increase in SE. Average SE was fairly similar between UIRT, augmented IRT, and MIRT for reading, speaking, and writing rubric domains. For the MC writing and listening domains, UIRT average SE was higher than for other domains, and closer to the CTT SE. The average SE decrease was less consistent for augmented IRT (e.g., the augmented IRT SE for speaking was higher than that for reading, but the reading reliability was lower than that of speaking).

Domain Subscore Reliability

If one was to rank the reliability of domains across grades, fairly consistently writing rubric had the highest reliability, followed by speaking for all methods except for augmented IRT (in which case it was reading), then followed by reading, MC writing, and listening. The least reliable domains (listening and MC writing) received the highest increase in reliability when augmented IRT and MIRT were used. For example, listening received approximately 21% increase in reliability from UIRT to augmented IRT on average across all grade bands, and the increase in reliability from UIRT to MIRT was 13%. For MC writing, the average increase in reliability after augmentation was 19%; the increase in reliability when MIRT was used was 12%. The initial CTT reliability of reading was fairly high, so the average increase in reliability

after augmentation was 5%; the average increase in reliability when MIRT was used was 2%. The initial CTT reliability of speaking was high, so the average increase in reliability after augmentation was only 1%; and average reliability decreased when MIRT was used. Writing rubric was consistently the domain with the highest reliability across all grades. The average increase in reliability after augmentation was only 1%; and average reliability decreased when MIRT was used. This finding was also supported by previous studies investigating subscore reporting by different methods (e.g., Wang, Chen, & Cheng, 2004; Dwyer et al., 2006; Wainer et al., 2001; Skorupski, 2008). Other factors that have been implicated in impacting reliability, such as the number of domains and test length, stayed the same (number of domains) or approximately the same (number of items, or the number of possible score points) in the examined assessment.

Differences in reliability between and within domains are only in part due to different methods of subscore estimation for those domains and varying correlations within pairs of those domains between grades. While only the method of subscore estimation was manipulated in this study, the reliability estimation of the subscores within a given domain for different grade bands is usually affected by a number of factors. It may be possible that constructs such as reading and writing, for example, is conceptualized differently for students in different grade bands. Consequently, it is likely that assessing proficiency in these domains is best accomplished with different, age-appropriate tasks across those grade bands. Due to these differences, it may be argued that domain reliability is not necessarily directly comparable within the same domain across grade bands. Similarly, since the meaning of the constructs changes across grade bands, it may be argued that reliability values across domains cannot be compared across grade bands. Additionally, a number of factors affect the reliability between language domains within the same subscore estimation method, even despite the fact that the number of possible points does

not differ drastically between domains. Of course, the differences between the number of students and the number of possible points does vary, as does the correlation between domains, which is important for those subscore estimation methods that capitalize on that correlation (i.e. the augmented IRT and MIRT). In addition, task-based assessments share common weaknesses pertaining to the nature of that type of assessment. One such weakness is the absence of a well-established framework for assessing the difficulty of different tasks (e.g. the task of "giving instructions" vs. the task of "providing an explanation"). Another weakness is a different interpretation of what the task entails by the examinees, depending on their characteristics, such as age, ability, or prior experience with similar assessments (Ellis, 2003). Additionally, limited research has been conducted to determine how different tasks affect performance on proficiency assessments. For example, Ejzenberg (2000) found that more fluent ELLs produced more discourse in uncued tasks (e.g. prompts to tell a story in which the speaker determines the structure of the story), while less fluent students produced the same amount of discource regardless of whether the task was cued (e.g., prompts to tell a story the structure of which is determined by the assessment administrator) or uncued. Poulisse and Schils (1989) found that higher-proficiency learners used more compensatory strategies when they did not know a word, and that the types of strategies they used differed according to the task in a way that was different from the lower-proficiency group. Generally, those studies that examine roles of different tasks are conducted using intermediate level ELLs (Skehan 2003). For that reason, it may be much

less clear how task types interact with proficiency levels other than the intermediate level.

In addition, construct-irrelevant variance that decreases reliability is introduced by some test characteristics. Language performance may vary along a number of dimensions according to the type of elicitation task which is used and the conditions under which it is implemented (Wigglesworth, 2000). For example, Bachman (1990) draws attention to a range of factors, characterized as trait facets and methods facets, which affect test performance and jeopardize test validity by introducing construct-irrelevant variance. Trait facets include such student characteristics as cognitive style and emotional status. Method facets, which have to do with the testing methods, include such factors as testing environment, test rubrics, the nature of the task, the nature of the expected response, and the interaction between the task and the response. In'nami and Koizumi (2009) proposed that among the many existing variables which affect language test scores, one central issue is the effect of task format on test performance. Item types and task types make an impact on the reliability and validity of the tests, with the tasks more appropriate to assess a specific construct producing more reliable and more defensible results. It may not always be clear, however, which tasks are the ones that are most appropriate to assess a specific construct. For example, in the assessment of the reading domain, Rupp, Ferne and Choi (2006) argue that multiple choice items are not a task that is naturally performed when one is reading a story. Riley and Lee (1996) found significant differences in the types of information given in summaries as opposed to recalls in reading comprehension tests. Additionally, content recall tasks normally used to assess reading may not be assessing the correct construct, but assess memory instead (Kobayashi, 2002). Text organization and test format had a significant impact on the students' performance in the reading domain. When texts used to test reading are clearly structured, the more proficient students achieved better results in summary writing and open-

ended questions. By contrast, the structure of the text made little difference to the performance of the less proficient students. This suggests that well-structured texts make it easier to differentiate between students with different levels of proficiency (Kobayashi, 2002).

Several authors draw attention to the fact that listening skills, while critical to language development, are not very well understood and hard to assess (Shin, 2008; Khoii & Paydarnia, 2001; Brown, 2004). We cannot observe the actual act of listening, nor can we see or hear an actual product of listening; instead, we can only observe the result of the test taker's auditory processing in the form of spoken or written responses. Thus, the assessment of listening can be done only by drawing inference from the test takers' speaking or writing in responding to aural passages. Hence, it is particularly important in listening tests to ensure that the questions actually measure the construct of listening comprehension (Shin, 2008). For example, multiple choice, a cloze format, or fill in the blank items of other forms may be used. According to Khoii and Paydarnia (2001), "a review of the literature on listening suggests that there is no generally accepted theory of listening comprehension on which to base these tests. It seems that in practice test constructors follow their instincts and just do their best when constructing tests of listening comprehension" (pp. 100-101). It may be due to this indeterminacy that listening domain has the lowest reliability on the state assessment examined in this study.

For a speaking assessment, some tasks, such as oral interviews have been praised for being more or less standardized. At the same time, this aspect reduces the authenticity of the assessment (Turner, 1998). According to Lantolf and Frawley (1988), an oral interview, being directed by the interviewer rather than allowing shared responsibility for the conversation, may not be an authentic form of assessing speaking. In addition, the reliance of most speaking assessments on raters introduces another aspect of construct-irrelevant variance. Studies have found that not only rater training and ability as an evaluator, but also differences in rater

behavior as a function of factors such as rater background, native language and amount of prior training or experience, impact the rater consistency in assessments of speaking (Upshur & Turner, 1999). While the speaking domain subscores were highly reliable in the state assessment examined, rater differences could have impacted the reliability, because this domain used tasks mentioned above assessed by individual raters.

A number of factors that arguably do not have a direct relationship to the evaluation of the construct of proficiency may affect the difficulty of the speaking and listening tasks, depending on the form of the task presentation, including the nature of the input (e.g., speech rate, length of passage, syntactic complexity, vocabulary, noise level, accent, amount of redundancy, etc.); the nature of the assessment task (e.g., the amount of context provided, clarity of instructions, response format, availability of question preview, etc.); and the individual listener factors (e.g., memory, interest, background knowledge, motivation, etc.). Features that have been found to affect the difficulty of listening (and other) comprehension items across a range of task types include amount of lexical overlap between the text and the response format; length of text preceding the information required to respond; length of required response; repetition of tested information; and whether responses and repetitions of information are verbatim or paraphrases (Brindley & Slatyer, 2002).

Some concerns related to the assessment of writing pertain to the authenticity of the writing tasks and the interrater reliability of assigning a score for a writing task. The issue with writing task authenticity is such that when high reliability is achieved by narrowing the range of task types or the range of skills tested, this restricts the interpretation of test results, and consequently the test validity (Saville, 2003). For example, if writing is tested with multiple-choice items, it may be more reliable, especially if the rater is taken completely out of the picture, but the authenticity of the task may be reduced. Similarly, isolating a specific task of

writing for testing purposes, such as testing only the correct spelling or punctuation, may increase reliability, but decrease authenticity, and consequently validity. In addition, it may not be clear to what authentic environments, if any, non-authentic task may generalize (Hawkey & Barker, 2004). Similar concerns may be voiced in regards to the tasks used by the writing domain of the state assessment examined.

Overall, it is likely that reliability for the same domain differed between grades because some tasks were more appropriate for a given grade than others, and because some concepts were easier to test at some grade bands than at others. The reliability differed between domains likely because tasks to assess some domains were more appropriate than tasks assessing other domains. Additionally, some tasks may be more reliable than others in assessing skills in general (such as, multiple choice tasks may be more reliable than essay constructed response tasks).

Relationship between Correlation and Reliability in Augmented IRT Subscores

CTT and UIRT do not use correlations between domains in subscore calculations. However, the augmented IRT and MIRT subscore calculations take into account these correlations. Augmented IRT scores are derived from unaugmented IRT scores. The higher the unaugmented IRT score correlation of the domain of interest with another domain, and the higher the unaugmented IRT score reliability of that other domain, the more impact that domain score had on the augmented domain score of interest.

A matrix of weights (the B matrix in Wainer et al.'s (2001) terminology) is constructed in the process of subscore augmentation, with the B values being the weights for the linear combination of the deviation scores that predict the best estimates for the subscale scores. The B weights indicate the magnitude of impact each domain is going to have on the score of the

domain of interest. The B matrix weights play a role similar to that of regression weights. The B weights are the highest for the domain that correlates most highly with the domain of interest, and has the highest reliability.

It is expected that the domain of interest itself would have the highest impact on its own score (Wainer et al., 2001), with the next highest impact being made by the domain it is most highly correlated with and most reliable. However, on several occasions, due to the reliability of the domain of interest being significantly lower than the reliability of other domains, the most impact was made by the domain with the highest correlation with the domain of interest and highest reliability.

For example, listening was consistently the least reliable domain across all grades and all methods, and consequently during subscore augmentation it received the most significant boost in reliability from other domains. For most grade bands, after listening itself, reading had the highest impact on the augmented listening subscore, because listening correlated most highly with reading, and reading was a highly reliable domain. However, in grade 0, speaking had the highest impact on augmented listening, which was even higher than the impact of listening itself. This impact magnitude was due to the fact that the reliability of listening was lower than the reliability of speaking, and listening and speaking correlated highly.

The unaugmented UIRT reliability of reading was fairly high compared to other domains. Therefore the reading subscore itself always had the highest impact on the augmented reading subscore. The MC writing subscore had the next highest impact on augmented reading, because it was reliable, and had the highest correlation with reading. On one occasion, however (in grade band 6-8), the reliability of writing rubric was higher than the reliability of reading; the

correlation of reading with writing rubric was high; therefore, writing rubric had a higher contribution to the augmented reading than reading itself.

Similarly, speaking was a domain with high reliability even prior to augmentation; it also did not have very high correlations with any other domain. Therefore, throughout all grade bands, speaking itself had the highest impact on the augmented speaking score, while other domains shared a small amount of impact on the speaking domain evenly.

In the MC writing domain, in grades 0 and 1, the reliability was substantively lower than that of reading, therefore reading had the biggest impact in those grades, even bigger than the MC writing score itself, on the augmented MC writing score for these grades. In grade band 4-5, writing and reading scores had an equal impact on the augmented writing score. Writing rubric was consistently the domain with the highest reliability across all grades, and therefore the writing rubric score itself had the highest impact on the writing rubric augmented score. The impact of other domains was fairly small, with the exception of a more significant reading impact on the writing rubric in grade band 6-8.

Correlation between Domains in MIRT-Estimated Subscores

MIRT estimates item parameters (slope and intercept) by direct estimation of the variance– covariance matrix of latent traits, and the loading of each item on the specified domain. In a confirmatory MIRT model, item parameters are calculated from the factor loadings and the estimated covariance matrix between dimensions, and then using these values, the student θ values were calculated. The higher the covariance between a specific latent trait and

the domain score of interest, the more impact that latent trait has on the augmented domain score of interest.

Generally, the pattern of latent trait covariance changes across grades and domains in MIRT was consistent with the pattern of domain correlation changes across grades and domains for augmented IRT subscores, with the average covariance across grades between listening and reading being .75, between listening and MC writing - .75, between listening and speaking - .61; between reading and MC writing - .88, between reading and speaking - .51, between speaking and MC writing - .53, between writing rubric and listening - .62, between writing rubric and reading - .67, between writing rubric and MC writing - .67, and between writing rubric and speaking - .61. Consequently, the impact of latent traits on each other in MIRT-derived subscores was similar to the pattern of impact of domains on each other in augmented IRT-derived subscores. For example, similar to augmented IRT, the MIRT-estimated listening trait was most highly and consistently impacted by the reading trait (in grade bands 2, 3, 4-5, 6-8); reading and MC writing shared an impact of similar magnitude on listening in grade band 9-12, and MC writing had the highest impact on listening in grades 0 and 1. The reading trait and the trait measured by MC writing items had the most high and consistent impact on each other. There was some variation in the degree and type of traits that had an impact on the MIRT-estimated speaking trait: listening had the highest impact in grades 0-3, and writing rubric – in grade bands 4-5, 6-8, and 9-12. However, the impact of all four other traits (listening, reading, MC writing, and writing rubric) was approximately the same on the speaking trait. Similar to speaking, all four other traits had an impact of similar magnitude on the writing rubric trait, with the MC writing having a somewhat higher impact in grades 2 and 3, MC writing and reading – in grade band 4-5, reading - in grade band 6-8, and speaking – in grade band 9-12.

Summary: Reliability of Different Methods of Subscore Assignment

Augmented IRT was found to be the most reliable method of subscore assignment out of the four methods we examined (CTT, UIRT, A-IRT, and MIRT). MIRT followed augmented IRT closely in reliability. MIRT standard error of measurement was the lowest of the four estimation methods, followed closely by the augmented IRT SE. CTT and UIRT, the methods that do not borrow information from other domains to estimate subscores, were the least reliable methods. In the process of augmentation, the least reliable domain (listening) received the biggest improvement in reliability, whereas in the domains that had high reliability prior to augmentation, the improvement in reliability was very small. Reliability generally increased, and SE generally decreased from grade 0 to grade band 9-12 for all domains.

Considerations for the Use of Augmented and Multidimensional Subscores

One factor that should be taken into consideration when augmenting subscores is whether the relationship between (unaugmented) domains is supported by prior research, and if so, to what degree. If that support is available, the augmentation of subscores and consequent increase in reliability is more defensible from the standpoint of subscore validity and interpretability. De la Torre and Patz (2005) advise that the valid use of scores obtained using auxiliary information of any type must be considered carefully. Even if all the information is obtained from the test performance itself, validity concerns remain. In addition, borrowing information from other domains may complicate the interpretation of augmented subscores due to the fact that after augmentation, these subscores contain information about other domains, as well as the domain of

interest. As mentioned previously, research on correlation between language domains is inconclusive, with relationships between domains varying from study to study. In situations where strong supporting evidence from research is not available about the relationship between domains, e.g., between listening and speaking, theoretical support for the augmentation of listening subscale with the speaking subscale, may not be readily available either.

Test developers and psychometricians are not always in the position to determine which different knowledge or ability domains get combined in the same test, but they can provide input regarding the effect that combining several different domains may have on the subscores, depending on the method of total score and subscore assignment. While in language testing, some support can be found for the relationship between all language domains, potentially even less related subjects or tasks can be combined on an assessment. Borrowing information from less related domains to enhance the reliability of the domain of interest may be less defensible from the standpoint of validity and interpretability in assessments in which less explicitly related domains are combined.

Additionally, if augmentation methods are used to estimate subscores that incorporate the reliability and correlation/ covariance between domains, for the scores to be invariant at different grade levels, these correlations between domains should be similar from one grade level to another. If the correlations between language domain scores differ substantially from one grade level to another, one can infer that the structure of the assessment is not the same across grade levels, and consequently the meaning of test scores is not equivalent across grade levels. While it is expected that correlations between domains change across grade levels and as proficiency develops, these differences may weaken the validity argument for a test that uses

augmented or multidimensional scores, as well as make the interpretability of the test results more complex.

An argument can be made that augmented subscores "artificially inflate" correlations between domains to boost reliability (Skorupski, 2008). Augmented IRT increases reliability of the domain of interest by making every student's profile more like the profiles of other students in the group, making it harder to differentiate between the abilities of individual students. The MIRT method, on the other hand, makes an individual student's score in one domain more like this same student's scores in other domains, making it harder to differentiate between the abilities of an individual student in different domains. It is the purpose of the test that usually dictates which of these methods may be more preferable (de la Torre, Song, & Hong, 2011; Wang, Chen, & Cheng, 2004). If the purpose of the test is to determine an individual student's strengths and weaknesses, the augmented IRT approach may be acceptable. If the purpose of the test is to determine how students differ from each other (e.g., for the purpose of giving out scholarships), augmented IRT may not be the best method. However, if the concern is the overall proficiency of an individual student, it may be legitimate to use the method that makes the student's scores in one domain more like the student's scores in other domains.

De la Torre, Song, and Hong (2011) suggest that augmented and multidimensional scores may not be desirable when straightforward interpretation of test scores as a summary of responses within a domain is favored. In any type of competition between students within a domain, it would not be appropriate to allow scores on other domains, especially not directly related ones, to affect the scores and ranking. However, when the consequences for examinees do not depend on comparisons with other examinee, then the greater accuracy of the multidimensional scores may be useful. For example, if more accurate score profiles lead to

more efficient diagnosis or more precise targeting of instructional resources, then their use could be supported.

Subscore Variability at Different Proficiency Levels

Examining variance and standard error of measurement at specific proficiency levels (also known as conditional standard error of measurement, or CSEM) provides information about the reliability of the test at those proficiency levels. If there are substantial differences in the variance and standard error of measurement at different levels of proficiency, the assessment may be considered less reliable at some proficiency levels than at others. In other words, if the assessment was hypothetically administered to a student of a given proficiency level several times, instead of being consistently placed in the same proficiency category after all those administrations, the student may be placed in different proficiency categories after some or all of those instances of administration. Which is more, we may be more or less certain about placing students in appropriate proficiency levels depending on their actual proficiency.

Overall, across domains and grades the standard deviation (SD) for all methods of subscore assignment becomes lower if the range of possible points for answering the assessment questions is narrow. In addition to some variation in the number of points possible between domains (e.g., there were 31 possible points in the speaking domain, but only 20 points in the writing rubric), the number of possible points varied from one proficiency level to another due to the cut scores set at those points. For example, for listening, the number of possible points at proficiency level 1 was on average 11.57; at level 2 – 5.57; at level 3 – 2.86, and at level 4 – 3.29. For speaking, the number of possible points at proficiency level 1 was on average 16; at proficiency level 2 – 6.86; at proficiency level 3 – 5.14; and at proficiency level 4, 4.57. For

reading, the number of possible points at proficiency level 1 was on average 10.71; at proficiency level 2 – 5.86; at proficiency level 3 - 4.00, and at proficiency level 4 – 5.14. Additionally, on average the percent of students at level 1 was 7%, at level 2 – 26%; at level 3 – 34%, and at level 4 – 33%. From grade band 3 on, level 4 had the highest percentage of students. Level 1 consistently had the lowest percentage of students across all grade levels.

Since the range of possible points was wide at proficiency level 1, the SD was consistently the highest at proficiency level 1. Then SD decreased at proficiency levels 2 and 3, because the range of possible points was narrow there. SD decreased at level 4 if there were fewer points available at that level than at level 3, and increased from level 3 in those instances when more points were available at level 4 than at level 3. MIRT had the lowest SD of the three IRT-derived subscores at all proficiency levels, followed by augmented IRT SD, and UIRT having the highest SD values at all proficiency levels. CTT-derived subscores consistently have the highest SD values of all four methods of subscore estimation at all proficiency levels. For example, for the listening domain, the average CTT SD for proficiency level 1 was 1.84; for level 2 – 1.44; for level 3 - .73, and for level 4 - .70. The average SD for UIRT proficiency level 1 was .64; for level 2 - .62; for level 3 - .60, and for level 4 - .55. For augmented IRT, the average SD for listening at proficiency level 1 was .55, for level 2 - .55, for level 3 - .51, and for level 4 - .48. For MIRT, the average SD for listening at proficiency level 1 was .47; for level 2 - .48; for level 3 - .45, and for level 4 - .51.

Subscore Standard Error of Measurement at Different Proficiency Levels

Similarly, MIRT had the lowest SE values of the three IRT-derived subscores at all proficiency levels, followed by augmented IRT SE, and UIRT having the highest SE values.

CTT-derived subscores consistently have the highest SE values of all four methods of subscore estimation at all proficiency levels. For example, for listening, the average SE for CTT subscores was 2.30 at proficiency level 1, 2.23 at proficiency level 2, 1.82 at proficiency level 3, and 1.13 at proficiency level 4. For UIRT, the average SE at proficiency level 1 was .45; at proficiency level 2, .51; at proficiency level 3, .60; and at proficiency level 4, .68. For MIRT, average SE for proficiency level 1 was .26; at proficiency level 2, .28; at proficiency level 3, .33, and at proficiency level 4, .38.

CSEM for CTT as measured by the formula suggested by Lord (1955) ($CSEm = \sqrt{[\frac{X(k-X)}{k-1}]}$) depends on how far the individual student's score is from the total number of points possible. For example, if a student's score is very close to, or is at, 0 or the maximum points possible, the CSEM for this individual would be 0. If the number of points possible in each proficiency level is equal, and there is no substantial disparity in the number of students at each score level, the intervals between average CSEM values for the four proficiency levels will likely be evenly distributed. However, more points were available at proficiency level 1 than at proficiency levels 2, 3, and 4. Additionally, there were more students at proficiency level 4 at all grade levels, and due to the point range restriction, they all had a score that was close to or at the maximum score possible. At the same time, there were very few students with a score of 0 or near 0 at any grade level or domain, even among those who were at proficiency level 1, as evidenced by the histograms of subscore distributions. It is likely due to this range restriction at specific proficiency levels and the skewed distribution of domain scores that the CTT CSEM was generally the highest at level 1, and either increasing slightly at level 2, then decreasing at levels 3 and 4, or decreasing from level 1 on. Similar to CTT CSEM values, due to the skewed

distribution of domain scores and the absence of subscores at the lowest level of proficiency, IRT CSEM generally was the lowest at proficiency level 1, and then increased to level 4.

Subscore Reliability at Different Proficiency Levels

Due to CTT subscore range restriction at specific proficiency levels, and consequently due to very limited variance of student (total) scores, as opposed to a more significant variance of item scores, very low or even negative values for CTT reliability (Cronbach's alpha) of subscores at specific proficiency levels resulted. Those values were substantially different from the reliability of domain subscores as a whole calculated for students at all levels of proficiency. Therefore those reliability values for specific proficiency levels were not compared with UIRT and MIRT reliability for specific proficiency levels In addition, since the reliability of augmented subscores is calculated as the ratio of the diagonal of the unconditional true score variance matrix to the diagonal of the unconditional estimated true score variance matrix, and is the average reliability for a specific domain, the reliability of augmented subscores at specific proficiency levels was not calculated.

UIRT and MIRT CSEM were estimated at different proficiency levels for grade bands and domains. The lower the SE for a specific proficiency level was, the higher the reliability for that proficiency level was as well. Since SE usually was the lowest at proficiency level 1, and increased to levels 2, 3, and 4, reliability generally was the highest at level 1 and decreased to levels 2, 3, and 4. MIRT reliability was generally higher at different proficiency levels than UIRT reliability. However, the reliability values for these two methods were considerably close together. Generally, IRT reliability values either remained close together for all proficiency levels, or decreased from level 1 to level 4 for both UIRT and MIRT. For

example, the average reliability for the listening domain estimated with UIRT was .63 for proficiency level 1, .56 for level 2, .47 for level 3, and .35 for level 4. The average reliability for the listening domain estimated with MIRT was .60 for proficiency level 1, .57 for level 2, .48 for level 3, and .49 for level 4.

Correlations between Domains at Different Proficiency Levels

Similar to evaluating the correlations between domains for the assessment of students at all proficiency levels, it is beneficial to evaluate those correlations for the scores of students at different proficiency levels. This evaluation allows to determine whether the structure of the assessment for students at different proficiency levels is similar, and the results of the assessment can be interpreted similarly for students at all proficiency levels, especially for subscore methods that borrow information from other correlated domains. However, correlations between domains at different proficiency levels did not appear to be consistent with correlations between domains before conditioning on proficiency level, with some correlations becoming very small or negative after the scores were grouped by total proficiency levels. What we encountered best fits the pattern of Berkson's paradox (Elwert & Winship, 2014; Westreich, 2012; Pearl, 2000), in which the assessment of the relationship between the domains becomes biased due to the fact that the total proficiency level was affected by proficiency in the individual domains.

Berkson's paradox, also known as endogenous selection bias, collider bias, collider-stratification bias, or collider-conditioning bias, is bias resulting from conditioning on a common effect of at least two causes. For example, factor A and factor B both cause outcome C. While collider bias may often occur when factors A and B are not related, it is also possible that there is a relationship between them even outside of conditioning on the outcome C. In real-life data,

variables that are completely unrelated can seldom be found (Westreich, 2012), which may complicate identification of collider bias. Conditioning on the collider may therefore introduce a correlation between the factors that are not related when not conditioned on the collider, or alter the relationship between two factors that are related in a different way when not conditioned on the collider.

Structural considerations are essential for assessing the impact of grouping variables in any specific way. It is the relational structure of the data that leads to bias when statistical analyses are performed with the data (Westreich, 2012). Specifically, conditioning on the common outcome of two (or more) variables (i.e., a collider) introduces a spurious association between them (Elwert & Winship, 2014). When factors A and B are truly not related when not conditioned on the collider, when that conditioning is present, it introduces the bias of these factors being related due to the causal effect of each on outcome C (collider variable).

When we attempted to perform correlations between domains after dividing the student population into proficiency levels, we were conditioning on the outcome (total proficiency level). While it was possible for a student to be at a specific total proficiency level, and not be at that proficiency level in individual domains, it was likely that the proficiency levels in individual domains were close to the overall proficiency level. Therefore the overall proficiency level was the outcome of the students being at specific domain proficiency levels, and was thus the collider variable. In addition, we assume based on previous linguistic research that the proficiency in one individual domains is related to the proficiency in other individual domains. This describes the situation in which the factors that are conditioned on the collider (overall proficiency level) are related. However, the correlation test performed to determine the degree to which they are related is biased due to conditioning on the collider.

Srivastava (2014) draws attention to another detail in the relationship between the factors and the collider variable: when the factors are in a compensatory relationship to the collider outcome, the correlation between the factors may become negative when conditioned on the collider, whereas without conditioning on the collider, the relationship between these factors may be positive. In our situation, it was possible for a student not to be at the same level of proficiency in all individual domains. It was also possible not to be at level 4 in all domains, and still be at level 4 for total proficiency. Since weights were applied in such a way that in grades 0 and 1 speaking and listening, in grades 3 and 4-5 reading and writing, and in bands 6-8 and 9-12 reading, writing, and listening could compensate for proficiency in other domains, negative correlations occurred between domains that were in that compensatory relationship when conditioned on the collider. Since reading and writing were the domains that could compensate for other domains, but never compensated for each other, the correlation between them never became negative when students were selected into groups based on their total proficiency level.

Elwert and Winship (2014) recommend several strategies to avoid collider bias. For example, they advise against conditioning on outcome variables when estimating total causal effects. They also stress the importance of examining the structure of the concept of interest and being aware of the underlying assumptions of statistical tests. Based on these suggestions, once we examined the data and realized that correlating the subscores after the population has been divided into ability levels is considered conditioning on the collider variable, we concluded that these correlations need to be evaluated with caution and likely do not reflect true correlations between these domains.

Summary: CSEM and Reliability at Different Proficiency Levels

Based on the information above regarding CSEM and reliability, we can conclude that we are more certain about the consistent placement of a student in proficiency level 1 than in proficiency levels 2, 3, or 4 based on IRT scores. We are also more certain about the consistent placement of a student in proficiency level 4 than in proficiency levels 1, 2, and 3 based on the CTT scores. Both UIRT and MIRT CSEM were lower and reliability was higher at proficiency level 1; both UIRT and MIRT CSEM were higher and reliability was lower at proficiency level 4. CTT CSEM was higher at proficiency level 1 and lower at proficiency level 4. However, it is challenging to make comparisons of reliability and precision at different levels of proficiency due to the fact that proficiency score intervals are not equal, and subscore distribution is skewed to the left. Generally, it is expected that IRT CSEM is the highest at the extremes of ability distribution, and the lowest – in the middle range of the ability distribution. While IRT CSEM was highest across all grades and domains at proficiency level 4, it was likely due not only to the fact that less information was available in determining the accurate placement of students on the ability continuum, but also due to the range restriction of available score points. In addition, IRT CSEM was generally lower at proficiency level 1 than at proficiency levels 2 and 3. This type of CSEM distribution may also be due the fact that ability in domains is not distributed normally across the population of ELL students, with a large number of students being at proficiency level 4 from grade band 3 on. Similarly, while CTT CSEM is expected to be the lowest at the extremes of ability distribution and highest in the middle of the distribution, it generally follows the expected pattern of being the lowest at proficiency level 4, but not at proficiency level 1. The low CTT CSEM at level 4 is due not only to more consistent responses of students at the extremes of proficiency, but also due to the range restriction of available score points at this level.

Future Research

While this study examined how each of the subscore assignment methods performed for individual proficiency levels, more research is needed that specifically examines how using multidimensional vs unidimensional scoring affects assignment to proficiency levels and classification accuracy. According to Ackerman, Gierl and Walker (2003), although a great deal of research has been conducted on the effects of model misspecification for item parameter estimation, little research has been conducted on the effects of model misspecification on ability parameter estimation. The few studies that have been done (e.g. Walker & Beretvas, 2003), found that incorrect references can be made about examinee proficiency when multidimensional data are analyzed with a unidimensional model, likely due to the fact that difficulty and dimensionality are confounded in the unidimensional estimate of ability, resulting in a multidimensional composite that does not remain consistent throughout the estimated unidimensional ability scale (Reckase, Carlson, Ackerman, & Spray, 1986).

It is important to perform dimensionality and reliability analyses on any test in circulation continuously. According to Ackerman, Gierl, and Walker (2003), "a complete test analysis should be viewed as an iterative craft, not just a straightforward application of measurement principles and formulas" (p.38). As we mentioned earlier, factor structure of a construct, and consequently of a test, may change from group to group and from year to year. This will affect how the students respond to the test, and their proficiency parameter estimation. The choice of scoring model in this possibly changing test environment will in turn affect the test reliability. Which is more, for some examinees, the response data may be unidimensional; for others, multidimensional. Thus, dimensionality analyses should be part of a standard set of analyses conducted after each test administration (Ackerman, Gierl, & Walker, 2003). Additionally, while

this study examines the impact of correlations between test dimensions, and thus indirectly, the test dimensionality, at the level of separate dimensions, the examination of the assessment dimensionality at the item level is needed as well. Specifically, as polytomous items are being used more frequently on large-scale assessments, the examination of the dimensionality of these polytomous items is necessary. How well the available multidimensional models for polytomously scored data represent the underlying structure in this response format has not yet been fully explored (Ackerman, Gierl, & Walker, 2003).

While factor analysis has been used to examine the dimensionality of the present assessment, some authors outline a number of problems with this approach. For example, while factor analysis assumes the linearity of the relationship between factors, it may not be such. Consequently, nonlinearity can result in a mismatch between the model and the data. Dimensionality can also be confounded with item difficulty, such that the factors represent items with comparable difficulty levels as opposed to items that measure distinct dimensions. Finally, there is often no agreement on how many factors should be viewed as meaningful; the number of factors identified in the course of factor analysis is often exploratory in nature and may not be supported by other research in the discipline related to the assessment content (Ackerman, Gierl, & Walker, 2003). Assessing dimensionality by fitting a MIRT model may be a better approach.

Another type of future research that can be conducted is comparing subscores based on simulated data based on the examinee responses and the actual examinee responses. Similar research can be conducted with data from several years of test administration to determine how correlations, factor loadings, test dimensionality, and subscore reliability compare to each other from year to year.

Another area of future research can be determining how the presence of polytomous items affects the use of different methods of subscore reporting. It is only recently that multidimensional models including polytomous items started being used; additional research is needed on how the use of these items affects the reliability and precision of subscore reporting. Additional models containing polytomous items need to be investigated in the MIRT environment.

In addition, models can be explored in which items are designed to measure several different traits. While in the unidimensional environment, we tend to avoid such items, they are a more natural representation of a construct. Between-items multidimensional tests have items in each test that are designed to measure the same latent trait, and different tests are designed to measure different latent traits; within-item multidimensional tests are tests in which some items are designed to simultaneously measure more than one single trait.

Limitations

One of the limitations of our study is that we used real data, which precluded us from knowing the true ability of the examinees. In addition, it precluded us from being able to separately examine the impact of such factors as student ability, age, number of students, number of subscales, correlation between subscales, and number of items in a subscale on the subscale reliability. Another limitation of using a real data set is that it is difficult to know how these results will generalize to other tests. For example, tests with different number of dimensions, different item types and formats, different correlational structures between domains may behave quite differently from what we observed in this study. We hope that they will generalize to other

tests of language proficiency, but it is unclear what the results may be for tests where the construct in question is structured differently.

Another limitation is that we had to use several different software programs to estimate subscores (e.g. SUBSKOR, flexMIRT). The use of different programs introduces additional variance in the subscore estimation. It may be helpful to either use one software throughout the project, or to compare the variability and reliability of subscores estimated by different software.

The limitation related to conclusions on CSEM was as follows: the proficiency levels we used to make comparisons between were assigned based on the cut scores for expected percent correct scores. They may or may not correspond closely to the "true" proficiency. There is as yet no absolute accepted cutoff point for advanced, intermediate, and beginner proficiency levels (Shin, 2005).

Conclusions and Policy Implications

Augmented IRT was found to be the most reliable method of subscore assignment out of the four methods we examined (CTT, UIRT, A-IRT, and MIRT). MIRT followed augmented IRT closely in reliability. MIRT standard error of measurement was the lowest of the four estimation methods, followed closely by the augmented IRT SE. CTT and UIRT, the methods that do not borrow information from other domains to estimate subscores, were the least reliable methods. In the process of augmentation, the least reliable domain (listening) received the biggest improvement in reliability, whereas in the domains that had high reliability prior to augmentation, the improvement in reliability was very small. Reliability generally increased, and SE generally decreased from grade 0 to grade band 9-12 for all domains.

Variability in domain subscores for all grade bands was the lowest for MIRT, followed closely by augmented IRT. Variability was higher in CTT and UIRT. Correlations between domain subscores were higher for MIRT and augmented IRT, and lower for CTT and UIRT. Correlations between domains generally increased from grade 0 to grade band 9-12.

Both UIRT and MIRT conditional standard errors of measurement were higher and reliability was lower at proficiency level 4 for most domains and grade bands. CTT CSEM was higher at proficiency level 1 and lower at proficiency level 4 for most grade bands. This is likely due to the fact that very few students were at the lowest level of proficiency. For those domains where some students were at the lowest level of proficiency (e.g., grade 0 speaking), CSEM for IRT-based methods was the highest and reliability was the lowest at levels 1 and 4; CSEM was the lowest and reliability was the highest at proficiency levels 2 and 3. Conversely, CTT CSEM was the highest and reliability was the lowest at proficiency levels 2 and 3, and CSEM was the lowest and reliability was the highest at proficiency levels 1 and 4.

When reliability of different methods is compared, the methods that borrow information from other domains (augmented IRT, MIRT) do have higher reliability and precision than those that do not make use of this information (CTT, UIRT). It is the purpose of the subscore use that should provide guidance as to what method of subscore assignment to use. Since subscores, and test scores in general, are seldom used for just one purpose, all these purposes may need to be taken into consideration in determining which subscore assignment methods are most appropriate .

De la Torre and Patz (2005) suggest that multidimensional scores may serve to complement, rather than to replace, traditional scoring of test batteries. Traditional unidimensional scores may be reported at the domain level (e.g., scale scores and their associated

norm- referenced and/or criterion-referenced derived scores), and multidimensional scores could be used to inform finer-grained reporting such as skills profiles and objective- level scores. While this suggestion may not be realistic for operational testing programs, which do not want to invest extra time and energy into an alternative scoring method that does not result in a reportable score, they may still want to compare different methods of subscore reporting and have a discussion about each method's reliability.

Wang, Chen and Cheng (2004) also caution that using collateral information might damage "face validity," because it is relatively difficult to explain to the public and test takers why a person's score on a test depends partly on his or her background or performance on other tests. Mislevy (1987) suggests that collateral information should be used very cautiously in estimates of persons' abilities when tests are used in individual selection or important placement decisions. In these situations, as much information that pertains to the construct directly should be gathered as possible, rather than relying on collateral information. In addition, for ease of interpretation, it is recommended that tests providing collateral information not be so divergent that one might question the reasonableness of adding information from them to a test taker's score. Consensus about the coverage of tests should be reached when very divergent tests are to be used as collateral information for important decisions. Wang, Chen and Cheng (2004) stress that although the improvement in measurement precision via the multidimensional approach is substantive, it does not mean that this approach should be applied to every test containing multiple domains. When reporting subscores, the decision on how many subscores to report should be guided by the usual and customary way of reporting scores and the ways of reporting that preserve test interpretability and test utility. If a test is split into several domains in a way in which the construct is not normally conceptualized, the interpretability and usefulness of such subscores may not be very high.

References

Abedi, J. (2004). The no child left behind act and English language learners: Assessment and accountability issues. *Educational Researcher*, *33*(1), 4-14.

Ackerman, T. A. (1994). Using multidimensional item response theory to understand what items and tests are measuring. *Applied Measurement in Education*, *7*(4), 255-278.

Adams, M. J. (1990). *Beginning to read: Thinking and learning about print*. Cambridge, Mass.

Alderson, J. (2007). The challenge of (diagnostic) testing: do we know what we are measuring? In J. Fox, M. Wesche, D. Bayliss, L. Cheng, C. Turner, & C. Doe (Eds.), *Language testing reconsidered* (pp.21-40). University of Ottawa Press: Ottawa, Ontario.

Alderson, J. C. (2005). *Assessing reading*. Ernst Klett Sprachen.

American Educational Research Association, American Psychological Association, National Council on Measurement in Education, Joint Committee on Standards for Educational, & Psychological Testing (US). (2014). *Standards for educational and psychological testing*. American Educational Research Association: New York.

August, D. E., & Shanahan, T. E. (2006). *Developing literacy in second-language learners: Report of the National Literacy Panel on Language-Minority Children and Youth*. Lawrence Erlbaum Associates Publishers.

Ayala, R. J., Schafer, W. D., & Sava-Bolesta, M. (1995). An investigation of the standard errors of expected a posteriori ability estimates. *British Journal of Mathematical and Statistical Psychology*, *48*(2), 385-405.

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

Bachman, L. F. (2007). The dialectic of abilities and contexts in defining constructs in language assessment. In J. Fox, M. Wesche, D. Bayliss, L. Cheng, C. Turner, & C. Doe (Eds.), *Language testing reconsidered* (pp.41-72). University of Ottawa Press: Ottawa, Ontario.

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

Bailey, A. L., & Butler, F. A. (2003). *An evidentiary framework for operationalizing academic language for broad application to K-12 education: A design document*. Center for the Study of Evaluation, National Center for Research on Evaluation, Standards, and Student Testing, Graduate School of Education & Information Studies, University of California, Los Angeles.

Berdie, R. F. (1969). Consistency and generalizability of intraindividual variability. *Journal of Applied Psychology*, *53*(1), 35.

Bock, R. D., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. *Psychometrika, 46,* 443-459.

Bock, R. D., & Mislevy, R. J. (1982). Adaptive EAP estimation of ability in a microcomputer environment. *Applied Psychological Measurement*, *6*(4), 431-444.

Bock, R. D., Thissen, D., & Zimowski, M. F. (1997). IRT estimation of domain scores. *Journal of Educational Measurement*, *34*(3), 197-211.

Boldt, R. F. (1988). *Latent structure analysis of TOEFL (TOEFL Research Report No. 20).* Princeton, NJ: Educational Testing Service.

Boughton, K. A., Yao, L., & Lewis, D. M. (2006). Reporting diagnostic subscale scores for tests composed of complex structure. In A*nnual Meeting of the National Council on Measurement in Education, San Francisco*.

Bozorgian, H. (2012). Metacognitive instruction does improve listening comprehension. *ISRN Education*, *2012*.

Brennan, R. L. (2012). *Utility indexes for decisions about subscores*. CASMA Research Report 33. Iowa City, IA: Center for Advanced Studies in Measurement and Assessment.

Brennan, R. L., & Lee, W. C. (1999). Conditional scale-score standard errors of measurement under binomial and compound binomial assumptions. *Educational and Psychological Measurement*, *59*(1), 5-24.

Brindley, G., & Slatyer, H. (2002). Exploring task difficulty in ESL listening assessment. *Language Testing*, *19*(4), 369-394.

Brown, H.D. (2004). *Language assessment: Principles and classroom practices*. New York: Longman.

Cai, L. (2010a). A two-tier full-information item factor analysis model with applications. *Psychometrika, 75,* 581–612.

Cai, L. (2010b). High-dimensional exploratory item factor analysis by a Metropolis-Hastings Robbins-Monro algorithm. *Psychometrika, 75,* 33– 57.

Cai, L. (2010c). Metropolis-Hastings Robbins-Monro algorithm for confirmatory item factor analysis. *Journal of Educational and Behavioral Statistics, 35,* 307–335.

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin*, *56*(2), 81.

Catts, H. W., Fey, M. E., Zhang, X., & Tomblin, J. B. (1999). Language basis of reading and reading disabilities: Evidence from a longitudinal investigation. *Scientific Studies of Reading*, *3*(4), 331-361.

Catts, H. W., Hogan, T. P., & Adlof, S. M. (2005). Developmental changes in reading and reading disabilities. In H.W. Catts, & A.G. Kamhi (Eds.), *The connections between language and reading disabilities* (pp. 25-40). Psychology Press.

Chall, J. S. (1983). *Learning to read: The great debate*. New York: McGraw-Hill.

Chalmers, R. P. (2012). Mirt: a multidimensional item response theory package for the R environment. *Journal of Statistical Software*, *48*(6), 1-29.

Chapelle, C. (2011). Conceptions of validity. In G. Fulcher and F. Davidson (Eds.), T*he Routledge handbook of language testing* (pp. 21-33). Routledge: London.

Chapelle, C. A., Chapelle, C. A., Enright, M. K., & Jamieson, J. M. (Eds.). (2011). *Building a validity argument for the Test of English as a Foreign Language:* Routledge.

Cheng, Y. Y., Wang, W. C., & Ho, Y. H. (2008). Multidimensional Rasch analysis of a psychological test with multiple subtests: A statistical solution for the bandwidth–fidelity dilemma. *Educational and Psychological Measurement*.

Chiappe, P., & Siegel, L. S. (2006). A longitudinal study of reading development of Canadian children from diverse linguistic backgrounds. *The Elementary School Journal*, *107*(2), 135-152.

Coleman, J. A. (2004). Modern languages in British universities past and present. *Arts and Humanities in Higher Education*, *3*(2), 147-162.

Crawford, J. (2004, September). No Child Left Behind: Misguided approach to school accountability for English language learners. In *Center on Educational Policy's Forum on Ideas to Improve the NCLB Accountability Provisions for Students with Disabilities and English Language Learners.*

Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory*. Holt, Rinehart and Winston: Orlando, FL.

Cronbach, L. J., Schönemann, P., & McKie, D. (1965). Alpha coefficients for Stratified-Parallel Tests. *Educational and Psychological Measurement*.

Culpepper, S. A. (2013). The reliability and precision of total scores and IRT estimates as a function of polytomous IRT parameters and latent trait distribution. *Applied Psychological Measurement*, *37*(3), 201-225.

Davis, C., & Bryant, P. (2006). Causal connections in the acquisition of an orthographic rule: a test of Uta Frith's developmental hypothesis. *Journal of Child Psychology and Psychiatry*, *47*(8), 849-856.

De Ayala, R. J. (2009). *Theory and practice of item response theory*. Guilford Publications.

de la Torre, J., & Patz, R. J. (2005). Making the most of what we have: A practical application of multidimensional item response theory in test scoring. *Journal of Educational and Behavioral Statistics*, *30*(3), 295-311.

de la Torre, J., & Song, H. (2009). Simultaneous estimation of overall and domain abilities: A higher-order IRT model approach. *Applied Psychological Measurement*.

de la Torre, J., Song, H., & Hong, Y. (2011). A comparison of four methods of IRT subscoring. *Applied Psychological Measurement*, *35*(4), 296-316.

DeMars, C. E. (2005). Scoring Subscales Using Multidimensional Item Response Theory Models. *Online Submission*.

Dimitrov, D. M. (2002). Reliability: Arguments for multiple perspectives and potential problems with generalization across studies. *Educational and Psychological Measurement*, *62*(5), 783-801.

Dimitrov, D. M. (2003). Marginal true-score measures and reliability for binary items as a function of their IRT parameters. *Applied Psychological Measurement*, *27*(6), 440-458.

Dunbar, S. B. (1982). Construct validity and the internal structure of a foreign language test for several native language groups. In *Annual Meeting of the American Educational Research Association, New York*.

Dwyer, A., Boughton, K. A., Yao, L., Steffen, M., & Lewis, D. (2006, April). *A comparison of subscale score augmentation methods using empirical data*. Paper presented at the meeting of the National Council on Measurement in Education, San Francisco, CA.

Edwards, M. C., & Vevea, J. L. (2006). An empirical Bayes approach to subscore augmentation: How much strength can we borrow? *Journal of Educational and Behavioral Statistics*, *31*(3), 241-259.

Ejzenberg, R. (2000). The juggling act of oral fluency: A psycho-sociolinguistic metaphor. In H. Riggenbach (Ed.), *Perspectives on fluency* (pp. 287–313). Ann Arbor: The University of Michigan Press.

Ellis, R. (2003). *Task-based language learning and teaching*. Oxford ; New York: OUP.

Elwert, F., & Winship, C. (2014). Endogenous Selection Bias: The Problem of Conditioning on a Collider Variable. *Annual Review of Sociology*, (0).

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. Psychology Press.

Farnia, F., & Geva, E. (2013). Growth and predictors of change in English language learners' reading comprehension. *Journal of Research in Reading*, *36*(4), 389-421.

Feldt, L. S. (1984). Some relationships between the binomial error model and classical test theory. *Educational and Psychological Measurement*, *44*(4), 883-891.

Feldt, L. S., & Brennan, R. L. (1989). Reliability. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 105-146). New York: American Council on Education and Macmillan.

Feldt, L. S., Steffen, M., & Gupta, N. C. (1985). A comparison of five methods for estimating the standard error of measurement at specific score levels. *Applied Psychological Measurement*, *9*(4), 351-361.

Finch, H. (2011). Multidimensional item response theory parameter estimation with nonsimple structure items. *Applied Psychological Measurement, 35*(1), 67-82.

Ford, K. L., Cabell, S. Q., Konold, T. R., Invernizzi, M., & Gartland, L. B. (2013). Diversity among Spanish-speaking English language learners: profiles of early literacy skills in kindergarten. *Reading and Writing*, *26*(6), 889-912.

Forrester, E. P. (2013). Longitudinal Relationships Between Reading and Spelling in Early Elementary Grades: Testing Causality Using a Cross-Lagged Panel Design.

Genesee, F., Lindholm-Leary, K., Saunders, W., & Christian, D. (2005). English language learners in US schools: An overview of research findings. *Journal of Education for Students Placed at Risk*, *10*(4), 363-385.

Gessaroli, M. E. (2004). Using hierarchical multidimensional item response theory to estimate augmented subscores. In A*nnual Meeting of the National Council on Measurement in Education, San Diego, CA*.

Geva, E. (2000). Issues in the assessment of reading disabilities in L2 children—beliefs and research evidence. *Dyslexia*, *6*(1), 13-28.

Geva, E. (2006). Learning to read in a second language: Research, implications, and recommendations for services. *Encyclopedia on Early Childhood Development, online resource. Montreal Quebec: Centre of Excellence for Early Childhood Development.*

Geva, E., & Farnia, F. (2012). Developmental changes in the nature of language proficiency and reading fluency paint a more complex view of reading comprehension in ELL and EL1. *Reading and Writing*, *25*(8), 1819-1845.

Gottardo, A. (2002). The relationship between language and reading skills in bilingual Spanish-English speakers. *Topics in Language Disorders*, *22*(5), 46-70.

Gottardo, A., & Mueller, J. (2009). Are first-and second-language factors related in predicting second-language reading comprehension? A study of Spanish-speaking children acquiring English as a second language from first to second grade. *Journal of Educational Psychology*, *101*(2), 330.

Green, B. F., Bock, R. D., Humphreys, L. G., Linn, R. L., & Reckase, M. D. (1984). Technical guidelines for assessing computerized adaptive tests. *Journal of Educational Measurement*, *21*(4), 347-360.

Griffin, P., Burns, M. S., & Snow, C. E. (Eds.). (1998). *Preventing reading difficulties in young children*. National Academies Press.

Gu, F., Little, T. D., & Kingston, N. M. (2013). Misestimation of reliability using coefficient alpha and structural equation modeling when assumptions of *tau* equivalence and uncorrelated errors are violated. *Methodology: European Journal of Research Methods for the Behavioral and Social Sciences*, *9*(1), 30.

Gulliksen, H. (1950). *Theory of mental tests.* Hillsdale, NJ: Lawrence Erlbaum Associates, Inc., Publishers.

Haberman, S. J. (2008). Subscores and validity. *ETS Research Report Series*, *2008*(2), i-11.

Haberman, S. J., & Sinharay, S. (2010). Reporting of subscores using multidimensional item response theory. *Psychometrika*, *75*(2), 209-227.

Haberman, S. J., & Sinharay, S. (2010). Reporting of subscores using multidimensional item response theory. *Psychometrika*, *75*(2), 209-227.

Haberman, S., Sinharay, S., & Puhan, G. (2009). Reporting subscores for institutions. *British Journal of Mathematical and Statistical Psychology*, *62*(1), 79-95.

Haertel, E. (2006). Reliability. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 65-110). Westport, CT: Praeger.

Hakuta, K., & August, D. (Eds.). (1997). *Improving Schooling for Language-Minority Children: A Research Agenda*. National Academies Press.

Hakuta, K., & Beatty, A. (2000). *Testing English-language learners in US schools*. Washington, DC: National Academy Press.

Haladyna, T. M., & Kramer, G. A. (2004). The validity of subscores for a credentialing test. *Evaluation & the Health Professions*, *27*(4), 349-368.

Hale, G. A. (1988). *Multiple-Choice Cloze Items and the Test of English as a Foreign Language. TOEFL Research Reports 26.*

Hale, G. A., Rock, D. A., & Jirele, T. (1982). Confirmatory Factor Analysis of the Test of English as a Foreign Language. *ETS Research Report Series*, *1982*(2), i-51.

Hale, G. A., Rock, D. A., & Jirele, T. (1989). *Confirmatory factor analysis of the Test of English as a ForeignLaugua (TOEFL Research Report No. 32).* Princeton, NJ: Educational Testing Service.

Hale, G. A., Stansfield, C. W., Rock, D. A., Hicks, M. M., Butler, F. A., & Oller, J. W. (1989). The relation of multiple-choice cloze items to the Test of English as a Foreign Language. *Language Testing*, *6*(1), 47-76.

Hambleton, R. K., & Plake, B. S. (1995). Using an extended Angoff procedure to set standards on complex performance assessments. *Applied Measurement in Education*, *8*(1), 41-55.

Hambleton, R. K., & Swaminathan, H. (1985). *Item response theory: Principles and applications* (Vol. 7). Springer.

Hambleton, R. K., Jones, R. W., & Rogers, H. J. (1993). Influence of item parameter estimation errors in test development. *Journal of Educational Measurement*, *30*(2), 143-155.

Hammill, D. D. (2004). What we know about correlates of reading. *Exceptional Children*, *70*(4), 453-469.

Hawkey, R., & Barker, F. (2004). Developing a common scale for the assessment of writing. *Assessing Writing*, *9*(2), 122-159.

He, Q. (2009). *Estimating the reliability of composite scores*. Coventry: Office of the Examinations and Qualifications Regulator (Ofqual).

Houts, C. R., & Cai, L. (2012). flexMIRT TM: Flexible Multilevel Item Factor Analysis and Test Scoring User's Manual Version 1.0.

Hoyt, C. (1941). Test reliability estimated by analysis of variance. *Psychometrika*, *6*(3), 153-160.

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

Jamieson, J., Jones, S., Kirsch, I., Mosenthal, P., & Taylor, C. (2000). *TOEFL 2000 framework*. Princeton, NJ: Educational Testing Service.

Jarjoura, D. (1986). An estimator of examinee-level measurement error variance that considers test form difficulty adjustments. *Applied Psychological Measurement*, *10*(2), 175-186.

Jones, N. (2011). Reliability and dependability. In G. Fulcher and F. Davidson (Eds.), T*he Routledge handbook of language testing* (pp. 350-361). Routledge: London.

Jongejan, W., Verhoeven, L., & Siegel, L. S. (2007). Predictors of reading and spelling abilities in first-and second-language learners. *Journal of Educational Psychology*, *99*(4), 835.

Kane, M. (2011). Articulating a validity argument. In G. Fulcher and F. Davidson (Eds.), T*he Routledge handbook of language testing* (pp. 34-47). Routledge: London.

Keats, J. A. (1957). Estimation of error variances of test scores. *Psychometrika*, *22*(1), 29-41.

Kelley, T. L. (1927). *Interpretation of educational measurements*. New York: World Book.

Kelley, T. L. (1947). *Fundamentals of statistics*. Harvard University Press.

Khoii, R., & Paydarnia, S. (2011). Test Method Facet and the Construct Validity of Listening Comprehension Tests. *The Journal of Applied Linguistics, 4(1),* 99-121.

Kieffer, M. J. (2011). Converging trajectories reading growth in language minority learners and their classmates, kindergarten to grade 8. *American Educational Research Journal*, *48*(5), 1187-1225.

Kieffer, M. J. (2012). Early oral language and later reading development in Spanish-speaking English language learners: Evidence from a nine-year longitudinal study. *Journal of Applied Developmental Psychology*, *33*(3), 146-157.

Kieffer, M. J., & Vukovic, R. K. (2013). Growth in reading-related skills of language minority learners and their classmates: More evidence for early identification and intervention. *Reading and Writing*, *26*(7), 1159-1194.

Kim, J. K., & Nicewander, W. A. (1993). Ability estimation for conventional tests. *Psychometrika*, *58*(4), 587-599.

Kim, S., & Feldt, L. S. (2010). The estimation of the IRT reliability coefficient and its lower and upper bounds, with comparisons to CTT reliability statistics. *Asia Pacific Education Review*, *11*(2), 179-188.

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

Kolen, M. J., Zeng, L., & Hanson, B. A. (1996). Conditional standard errors of measurement for scale scores using IRT. *Journal of Educational Measurement*, *33*(2), 129-140.

Kunnan, A. (2011). Language assessment for immigration and citizenship. In G. Fulcher and F. Davidson (Eds.), T*he Routledge handbook of language testing* (pp. 162-177). Routledge: London.

Lantolf, J. P., & Frawley, W. (1988). Proficiency. *Studies in Second Language Acquisition*, *10*(2), 181-195.

Lee, W. C., Brennan, R. L., & Kolen, M. J. (2000). Estimators of Conditional Scale-Score Standard Errors of Measurement: A Simulation Study. *Journal of Educational Measurement*, *37*(1), 1-20.

Linacre, J. M. (1997). KR-20 or Rasch reliability: Which tells the "truth". *Rasch Measurement Transactions*, *11*(3), 580-581.

Lindsey, K. A., Manis, F. R., & Bailey, C. E. (2003). Prediction of first-grade reading in Spanish-speaking English-language learners. *Journal of Educational Psychology*, *95*(3), 482.

Ling, G. (2009, April). Why the major field (business) test does not report subscores of individual test-takers—reliability and construct validity evidence. In *Annual Meeting of the National Council on Measurement in Education, San Diego, CA*.

Lipka, O., & Siegel, L. S. (2007). The development of reading skills in children with English as a second language. *Scientific Studies of Reading*, *11*(2), 105-131.

Liu, J., & Costanzo, K. (2013). The relationship among TOEIC listening, reading, speaking, and writing skills. *The research foundation for the TOEIC tests: A compendium of studies*, *2*, 2-1.

Lonigan, C. J., Schatschneider, C., & Westberg, L. The National Early Literacy Panel.(2008). Identification of children's skills and abilities linked to later outcomes in reading, writing, and spelling. *National Institute for Literacy (Ed.), Developing early literacy: Report of the national early literacy panel*, 55-106.

Lord, F. M. (1955). Estimating test reliability. *Educational and Psychological Measurement*, 15, 325-336.

Lord, F. M. (1984). Standard errors of measurement at different ability levels. *Journal of Educational Measurement*, *21*(3), 239-243.

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.

Lord, F.M. (1965). A strong true score theory, with applications. *Psychometrika*, *30,* 239-270.

Lord, F.M. (1980). *Applications of item response theory to practical testing problems*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Luecht, R. M. (2003). *Applications of Multidimensional Diagnostic Scoring for Certification and Licensure Tests.* Paper presented at the Annual Meeting of the National Council on Measurement in Education (Chicago, IL, April 21-25, 2003).

Mancilla-Martinez, J., & Lesaux, N. K. (2010). Predictors of reading comprehension for struggling readers: The case of Spanish-speaking language minority learners. *Journal of Educational Psychology*, *102*(3), 701.

Mancilla-Martinez, J., Kieffer, M. J., Biancarosa, G., Christodoulou, J. A., & Snow, C. E. (2011). Investigating English reading comprehension growth in adolescent language minority learners: Some insights from the simple view. *Reading and Writing*, *24*(3), 339-354.

Manis, F. R., Lindsey, K. A., & Bailey, C. E. (2004). Development of Reading in Grades K–2 in Spanish-Speaking English-Language Learners. *Learning Disabilities Research & Practice*, *19*(4), 214-224.

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika*, *47*(2), 149-174.

McKinley, R. L., & Way, W. D. (1992). The Feasibility of Modeling Secondary TOEFL Ability Dimensions Using Multidimensional TRT Models.

McNamara, T., & Roever, C. (2006). *Language testing: The social dimension.* (Vol. 1). John Wiley & Sons.

Messick, S.A. (1981). Constructs and their vicissitudes in educational and psychological measurement. *American Psychologist, 89,* 575-588.

Mislevy, R. J. (1987). Exploiting auxiliary information about examinees in the estimation of item parameters. *Applied Psychological Measurement*, *11*(1), 81-91.

Mislevy, R. J., & Stocking, M. L. (1989). A consumer's guide to LOGIST and BILOG. *Applied Psychological Measurement*, *13*(1), 57-75.

Mislevy, R. J., Beaton, A. E., Kaplan, B., & Sheehan, K. M. (1992). Estimating population characteristics from sparse matrix samples of item responses. *Journal of Educational Measurement*, *29*(2), 133-161.

Mislevy, R., and Yin, C. (2011). Evidence-centered design in language testing. In G. Fulcher and F. Davidson (Eds.), T*he Routledge handbook of language testing* (pp. 208-222). Routledge: London.

Mislevy, R. J. (2009). Validity from the perspective of model-based reasoning. *The concept of validity: Revisions, new directions, and applications*, 83-108.

Mollenkopf, W. (1949). Variation of the standard error of measurement. *Psychometrika*, 154, 189-229.

Morris, D., Bloodgood, J., & Perney, J. (2003). Kindergarten predictors of first-and second-grade reading achievement. *The Elementary School Journal*, 93-109.

Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *Applied Psychological Measurement*, *16*(2), 159-176.

Muthén, L. K., & Muthén, B. O. (2008). Mplus (Version 5.1). *Los Angeles, CA: Muthén & Muthén*.

Nakamoto, J., Lindsey, K. A., & Manis, F. R. (2007). A longitudinal analysis of English language learners' word decoding and reading comprehension. *Reading and Writing*, *20*(7), 691-719.

Nandakumar, R. (1994). Assessing dimensionality of a set of item responses - comparison of different approaches. *Journal of Educational Measurement*, *31*(1), 17-35.

Oltman, P. K., Stricker, L. J., & Barrows, T. S. (1988). *Native language, English proficiency, and the structure of the Test of English as a Foreign Language*. Educational Testing Service.

Pearl, J. (2000). *Causality: models, reasoning and inference* (Vol. 29). Cambridge: MIT press.

Peyton, V., Kingston, N.M., Skorupski, W., Glasnapp, D., & Poggio, J. (2009). *State English language proficiency assessment technical manual.* Center for Educational Testing and Evaluation, University of State.

Powers, D. E. (2010). The case for a comprehensive, four-skills assessment of English-language proficiency. *R & D Connections* (14).

Powers, D. E. (2013). Assessing English language proficiency in all four language domains: Is it really necessary? *The research foundation for the TOEIC tests: A compendium study.*

Puhan, G., Sinharay, S., Haberman, S., & Larkin, K. (2010). The utility of augmented subscores in a licensure exam: An evaluation of methods using empirical data. *Applied Measurement in Education*, *23*(3), 266-285.

Raju, N. S., Price, L. R., Oshima, T. C., & Nering, M. L. (2007). Standardized conditional SEM: A case for conditional reliability. *Applied Psychological Measurement*, *31*(3), 169-180.

RAND Reading Study Group. (2002). *Reading for understanding: Towards an R&D program in reading comprehension.* Retrieved October 1, 2014, from http://www.rand.org/multi/achievementforall/reading/readreport.html

Rawls, J. (Ed.). (2001). *Justice as fairness: A restatement*. Harvard University Press.

Reckase, M. D. (1985). The difficulty of test items that measure more than one ability. *Applied Psychological Measurement*, *9*(4), 401-412.

Reckase, M. D. (2009). *Multidimensional item response theory*. New York: Springer.

Reckase, M. D., & Hirsch, T. M. (1991). *Interpretation of Number-Correct Scores when the True Number of Dimensions Assessed by a Test Is Greater than Two.* Paper presented at the Annual Meeting of the National Council on Measurement in Education (Chicago, IL, April 4, 1991).

Riley, G. L., & Lee, J. F. (1996). A comparison of recall and summary protocols as measures of second language reading comprehension. *Language Testing*, *13*(2), 173-189.

Ross, S. (2011). Claims, evidence, and inference in performance assessment. In G. Fulcher and F. Davidson (Eds.), T*he Routledge handbook of language testing* (pp. 223-234). Routledge: London.

Rupp, A. A., Ferne, T., & Choi, H. (2006). How assessing reading comprehension with multiple-choice questions shapes the construct: A cognitive processing perspective. *Language Testing*, *23*(4), 441-474.

Samejima, F. (1977). A use of the information function in tailored testing. *Applied Psychological Measurement, 1*(2), 233-247.

Saville, N (2003). The process of test development and revision within Cambridge EFL. In Weir, C., and Milanovic, M. (Eds), *Continuity and Innovation: Revising the Cambridge Proficiency in English Examination 1913–2002*. Cambridge: Cambridge ESOL/Cambridge.

Sawaki, Y., Stricker, L. J., & Oranje, A. H. (2009). Factor structure of the TOEFL Internet-based test. *Language Testing*, *26*(1), 005-30.

Sawaki, Y., Stricker, L., & Oranje, A. (2008). *Factor structure of the TOEFL Internet-based test (iBT): Exploration in a field trial sample.* Educational Testing Service, Princeton, NJ.

Scarborough, H. S. (1998). Predicting the future achievement of second graders with reading disabilities: Contributions of phonemic awareness, verbal memory, rapid naming, and IQ. *Annals of Dyslexia*, *48*(1), 115-136.

Schatschneider, C., Fletcher, J. M., Francis, D. J., Carlson, C. D., & Foorman, B. R. (2004). Kindergarten prediction of reading skills: A longitudinal comparative analysis. *Journal of Educational Psychology*, *96*(2), 265.

Shanahan, T., & Lomax, R. G. (1986). An analysis and comparison of theoretical models of the reading–writing relationship. *Journal of Educational Psychology*, *78*(2), 116.

Shanahan, T., & Lomax, R. G. (1988). A developmental comparison of three theoretical models of the reading-writing relationship. *Research in the Teaching of English*, 196-212.

Shin, D. (2007). *A comparison of methods of estimating subscale scores for mixed-format tests*. Pearson Educational Measurement.

Shin, S. K. (2005). Did they take the same test? Examinee language proficiency and the structure of language tests. *Language Testing*, *22*(1), 31-57.

Shin, S. K. (2005). Did they take the same test? Examinee language proficiency and the structure of language tests. *Language Testing*, *22*(1), 31-57.

Shin, S. Y. (2008). *Examining the construct validity of a web-based academic listening test: an investigation of the effects of response formats. SPAAN FELLOW*, *1001*, 95.

Shohamy, E. (2006). *Language policy: hidden agendas and new approaches.* Psychology Press.

Shu, L., & Schwarz, R. D. (2014). IRT-Estimated Reliability for Tests Containing Mixed Item Formats. *Journal of Educational Measurement*, *51*(2), 163-177.

Sinharay, S. (2010). How often do subscores have added value? Results from operational and simulated data. *Journal of Educational Measurement*, *47*(2), 150-174.

Sinharay, S., & Haberman, S. J. (2008). Reporting subscores: A survey. *Research Memorandum*, (08-18).

Sinharay, S., Haberman, S., & Puhan, G. (2007). Subscores based on classical test theory: To report or not to report. *Educational Measurement: Issues and Practice*, *26*(4), 21-28.

Sinharay, S., Puhan, G., & Haberman, S. J. (2010). Reporting diagnostic scores in educational testing: Temptations, pitfalls, and some solutions. *Multivariate Behavioral Research*, *45*(3), 553-573.

Sinharay, S., Puhan, G., & Haberman, S. J. (2011). An NCME instructional module on subscores. *Educational Measurement: Issues and Practice*, *30*(3), 29-40.

Skehan, P. (2003). Task-based instruction. *Language Teaching, 36*(1), 1-14.

Skorupski, W. P. (2008, August). A review and empirical comparison of approaches for improving the reliability of objective level scores. In *annual meeting of A Study of the Council of Chief State School Officers*.

Skorupski, W. P., & Carvajal, J. (2009). A comparison of approaches for improving the reliability of objective level scores. *Educational and Psychological Measurement*.

Solano-Flores, G., & Trumbull, E. (2003). Examining language in context: The need for new research and practice paradigms in the testing of English-language learners. *Educational Researcher*, *32*(2), 3-13.

Spearman, C. (1904). The proof and measurement of association between two things. *The American Journal of Psychology*, *15*(1), 72-101.

Spolsky, B. (2007). On second thoughts. In J. Fox, M. Wesche, D. Bayliuss, L. Cheng, C. Turner, & C. Doe (Eds.), *Language testing reconsidered* (pp.9-21). University of Ottawa Press: Ottawa, Ontario.

Srivastava, S. (2014, August 4). The selection-distortion effect: How selection changes correlations in surprising ways. Retrieved from

https://hardsci.wordpress.com/2014/08/04/the-selection-distortion-effect-how-selection-changes-correlations-in-surprising-ways/

Stanovich, K. E. (1993). The language code: Issues in word recognition. In *Reading across the life span* (pp. 111-135). Springer New York.

Stone, C. A., Ye, F., Zhu, X., & Lane, S. (2009). Providing subscale scores for diagnostic information: A case study when the test is essentially unidimensional. *Applied Measurement in Education*, *23*(1), 63-86.

Storch, S. A., & Whitehurst, G. J. (2002). Oral language and code-related precursors to reading: evidence from a longitudinal structural model. *Developmental Psychology*, *38*(6), 934.

Stricker, L. J., Rock, D. A., & Lee, Y. W. (2005). *Factor structure of the languEdge™ test across language groups. Research report #5.* Princeton, NJ: Educational Testing Service.

Swinton, S. S., & Powers, D. E. (1980). *Factor analysis: the Test of English as a Foreign Language. TOEFL Research Report No. 6)*. Princeton, NJ: Educational Testing Service.

Thissen, D. (2000). Reliability and measurement precision. In H. Wainer (Ed). *Computerized adaptive testing: A primer* (2nd ed.), pp. 159-184. Mahwah, NJ, US: Lawrence Erlbaum Associates Publishers.

Thissen, D., & Wainer, H. (Eds). *Test scoring.* Mahwah, NJ, US: Lawrence Erlbaum Associates Publishers.

Thissen, D., & Edwards, M. C. (2005). Diagnostic scores augmented using multidimensional item response theory: Preliminary investigation of MCMC strategies. In *annual meeting of the National Council on Measurement in Education, Montreal, CA*.

Thissen, D., & Orlando, M. (2001). Item response theory for items scored in two categories. In D. Thissen & H. Wainer (Eds.), *Test Scoring* (pp.73-140). Mahwah, NJ, US: Lawrence Erlbaum Associates Publishers.

Thissen, D., Cai, L., & Bock, R.D. (2010). The nominal categories item response model. In M. Nering & R. Ostini (Eds.), *Handbook of polytomous item response theory models* (pp.43-75). New York, NY: Routledge.

Thorndike, R. L. (1951). Community variables as predictors of intelligence and academic achievement. *Journal of Educational Psychology*, *42*(6), 321.

Turner, J. (1998). Assessing speaking. *Annual Review of Applied Linguistics*, *18*, 192-207.

Upshur, J.A. and Turner, C.E. 1999: Systematic effects in the rating of second language speaking ability: test method and learner discourse. *Language Testing* 16, 82–111.

Vellutino, F. R., Scanlon, D. M., Small, S. G., & Tanzman, M. S. (1991). The linguistic basis of reading ability: Converting written to oral language. *Text, 11,* 99–133.

Verhoeven, L., & van Leeuwe, J. (2012). The simple view of second language reading throughout the primary grades. *Reading and Writing*, *25*(8), 1805-1818.

Wagner, R. K., Torgesen, J. K., Rashotte, C. A., Hecht, S. A., Barker, T. A., Burgess, S. R., ... & Garon, T. (1997). Changing relations between phonological processing abilities and word-level reading as children develop from beginning to skilled readers: a 5-year longitudinal study. *Developmental Psychology*, *33*(3), 468.

Wainer, H., Vevea, J. L., Camacho, F., Reeve, B. B., Rosa, K., Nelson, L., ... & Thissen, D. (2001). Augmented scores—"borrowing strength" to compute scores based on small numbers of items. In D. Thissen & H. Wainer (Eds.), *Test scoring* (pp. 343-387). Routledge.

Walker, C.M., & Beretvas, S. N. (2003). Comparing multidimensional and unidimensional proficiency classifications: Multidimensional IRT as a diagnostic aid. *Journal of Educational Measurement*, *40*(3), 255-275.

Wang, W. C., Chen, P. H., & Cheng, Y. Y. (2004). Improving measurement precision of test batteries using multidimensional item response models. *Psychological Methods*, *9*(1), 116.

Warley, H. P., Landrum, T. J., Invernizzi, M. A., & Justice, L. (2005). *Prediction of first grade reading achievement: A comparison of kindergarten predictors.* In Yearbook – National Reading Conference (Vol. 54, p. 428).

Westreich, D. (2012). Berkson's bias, selection bias, and missing data. *Epidemiology (Cambridge, Mass.)*, *23*(1), 159.

Whitehurst, G. J., & Lonigan, C. J. (1998). Child development and emergent literacy. *Child Development*, *69*(3), 848-872.

Wigglesworth, G. (2000). Issues in the development of oral tasks for competency-based assessments of second language performance. *Studies in Immigrant English Language Assessment*, *1*, 81-124.

Wolf, M. K., Herman, J. L., Kim, J., Abedi, J., Leon, S., Griffin, N., ... & Shin, H. W. (2008). Providing Validity Evidence to Improve the Assessment of English Language Learners. CRESST Report 738. *National Center for Research on Evaluation, Standards, and Student Testing (CRESST)*.

Woodruff, D. (1990). Conditional standard error of measurement in prediction.*Journal of Educational Measurement*, *27*(3), 191-208.

Xu, T., & Stone, C. A. (2012). Using IRT trait estimates versus summated scores in predicting outcomes. *Educational and Psychological Measurement*, *72*(3), 453-468.

Yao, L. (2003). BMIRT: Bayesian multivariate item response theory. *[Computer software]*. *Monterey, CA: CTB/McGraw-Hill*.

Yao, L. (2010). Reporting valid and reliable overall scores and domain scores. *Journal of Educational Measurement*, *47*(3), 339-360.

Yao, L., & Boughton, K. A. (2007). A multidimensional item response modeling approach for improving subscale proficiency estimation and classification. *Applied Psychological Measurement*, *31*(2), 83-105.

Yao, L., & Schwarz, R. D. (2006). A multidimensional partial credit model with associated item and test statistics: An application to mixed-format tests. *Applied Psychological Measurement*, *30*(6), 469-492.

Yen, W. M. (1987, June). A Bayesian/IRT index of objective performance. In A*nnual Meeting of the Psychometric Society, Montreal, Quebec, Canada*.

Appendix 1

	Item	λ1	SE λ1	λ2	SE λ2	λ3	SE λ3	λ4	SE λ4
Listening	1	0.53	0.05	0	----	0	----	0	----
	2	0.45	0.04	0	----	0	----	0	----
	3	-0.5	0.05	0	----	0	----	0	----
	4	0.59	0.03	0	----	0	----	0	----
	5	0.52	0.03	0	----	0	----	0	----
	6	0.48	0.03	0	----	0	----	0	----
	7	0.39	0.06	0	----	0	----	0	----
	8	0.47	0.03	0	----	0	----	0	----
	9	0.41	0.03	0	----	0	----	0	----
	10	0.54	0.03	0	----	0	----	0	----
	11	0.38	0.04	0	----	0	----	0	----
	12	0.23	0.04	0	----	0	----	0	----
	13	0.38	0.04	0	----	0	----	0	----
	14	0.37	0.04	0	----	0	----	0	----
	15	0.17	0.03	0	----	0	----	0	----
	16	0.34	0.03	0	----	0	----	0	----
	17	0.38	0.03	0	----	0	----	0	----
	18	0.5	0.03	0	----	0	----	0	----
Reading	19	0	----	0.74	0.07	0	----	0	----
	20	0	----	0.82	0.05	0	----	0	----
	21	0	----	0.82	0.05	0	----	0	----
	22	0	----	0.69	0.04	0	----	0	----
	23	0	----	0.7	0.04	0	----	0	----
	24	0	----	0.72	0.03	0	----	0	----
	25	0	----	0.67	0.03	0	----	0	----
	26	0	----	0.88	0.02	0	----	0	----
	27	0	----	0.84	0.02	0	----	0	----
	28	0	----	0.73	0.02	0	----	0	----
	29	0	----	0.85	0.02	0	----	0	----
	30	0	----	0.8	0.02	0	----	0	----
	31	0	----	0.89	0.01	0	----	0	----
	32	0	----	0.39	0.04	0	----	0	----
	33	0	----	0.69	0.02	0	----	0	----
	34	0	----	0.43	0.03	0	----	0	----
	35	0	----	0.72	0.02	0	----	0	----
	36	0	----	0.78	0.02	0	----	0	----
	37	0	----	0.75	0.02	0	----	0	----
	38	0	----	0.77	0.02	0	----	0	----
	39	0	----	0.78	0.02	0	----	0	----

	40	0	----	0.88	0.01	0	----		0	----				
	41	0	----	0.78	0.02	0	----		0	----				
	42	0	----	0.55	0.03	0	----		0	----				
Writing	43	0	----	0	----	0.57	0.03		0	----				
	44	0	----	0	----	0.43	0.05		0	----				
	45	0	----	0	----	0.51	0.07		0	----				
	46	0	----	0	----	0.55	0.03		0	----				
	47	0	----	0	----	0.62	0.05		0	----				
	48	0	----	0	----	0.5	0.06		0	----				
	49	0	----	0	----	0.48	0.04		0	----				
	50	0	----	0	----	0.53	0.03		0	----				
	51	0	----	0	----	0.57	0.03		0	----				
	52	0	----	0	----	0.65	0.03		0	----				
	53	0	----	0	----	0.71	0.02		0	----				
	54	0	----	0	----	0.54	0.03		0	----				
	55	0	----	0	----	0.75	0.02		0	----				
	56	0	----	0	----	0.5	0.03		0	----				
Speaking	57	0	----	0	----	0	----		0.54	0.03				
	58	0	----	0	----	0	----		0.46	0.02				
	59	0	----	0	----	0	----		0.65	0.02				
	60	0	----	0	----	0	----		0.7	0.02				
	61	0	----	0	----	0	----		0.69	0.02				
	62	0	----	0	----	0	----		0.67	0.02				
	63	0	----	0	----	0	----		0.79	0.02				
	64	0	----	0	----	0	----		0.82	0.01				
	65	0	----	0	----	0	----		0.81	0.01				

Table 1. Grade 1 item factor loadings.

	Item	a 1	s.e.	a 2	s.e.	a 3	s.e.	a 4	s.e.
Listening	1	1.05	0.09	0	----	0	----	0	----
	2	0.85	0.05	0	----	0	----	0	----
	3	-0.98	0.07	0	----	0	----	0	----
	4	1.24	0.07	0	----	0	----	0	----
	5	1.04	0.05	0	----	0	----	0	----
	6	0.94	0.05	0	----	0	----	0	----
	7	0.72	0.07	0	----	0	----	0	----
	8	0.9	0.05	0	----	0	----	0	----
	9	0.78	0.04	0	----	0	----	0	----
	10	1.1	0.05	0	----	0	----	0	----
	11	0.69	0.05	0	----	0	----	0	----
	12	0.4	0.04	0	----	0	----	0	----
	13	0.7	0.05	0	----	0	----	0	----
	14	0.68	0.04	0	----	0	----	0	----
	15	0.29	0.04	0	----	0	----	0	----
	16	0.62	0.04	0	----	0	----	0	----
	17	0.7	0.04	0	----	0	----	0	----
	18	0.98	0.04	0	----	0	----	0	----
Reading	19	0	----	1.87	0.22	0	----	0	----
	20	0	----	2.43	0.26	0	----	0	----
	21	0	----	2.48	0.25	0	----	0	----
	22	0	----	1.61	0.11	0	----	0	----
	23	0	----	1.69	0.1	0	----	0	----
	24	0	----	1.75	0.08	0	----	0	----
	25	0	----	1.53	0.08	0	----	0	----
	26	0	----	3.12	0.17	0	----	0	----
	27	0	----	2.65	0.13	0	----	0	----
	28	0	----	1.84	0.08	0	----	0	----
	29	0	----	2.79	0.12	0	----	0	----
	30	0	----	2.3	0.09	0	----	0	----
	31	0	----	3.27	0.13	0	----	0	----
	32	0	----	0.72	0.05	0	----	0	----
	33	0	----	1.62	0.06	0	----	0	----
	34	0	----	0.82	0.04	0	----	0	----
	35	0	----	1.76	0.06	0	----	0	----
	36	0	----	2.09	0.1	0	----	0	----
	37	0	----	1.91	0.08	0	----	0	----
	38	0	----	2.04	0.08	0	----	0	----
	39	0	----	2.11	0.07	0	----	0	----
	40	0	----	3.13	0.12	0	----	0	----
	41	0	----	2.14	0.09	0	----	0	----
	42	0	----	1.11	0.05	0	----	0	----

Writing	43	0	----	0	----	1.19	0.06	0	----
	44	0	----	0	----	0.81	0.07	0	----
	45	0	----	0	----	1.02	0.11	0	----
	46	0	----	0	----	1.11	0.05	0	----
	47	0	----	0	----	1.33	0.11	0	----
	48	0	----	0	----	0.97	0.09	0	----
	49	0	----	0	----	0.93	0.05	0	----
	50	0	----	0	----	1.08	0.05	0	----
	51	0	----	0	----	1.17	0.05	0	----
	52	0	----	0	----	1.46	0.06	0	----
	53	0	----	0	----	1.71	0.07	0	----
	54	0	----	0	----	1.1	0.05	0	----
	55	0	----	0	----	1.93	0.08	0	----
	56	0	----	0	----	0.99	0.05	0	----
Speaking	57	0	----	0	----	0	----	1.08	0.04
	58	0	----	0	----	0	----	0.88	0.04
	59	0	----	0	----	0	----	1.46	0.05
	60	0	----	0	----	0	----	1.68	0.06
	61	0	----	0	----	0	----	1.63	0.05
	62	0	----	0	----	0	----	1.53	0.05
	63	0	----	0	----	0	----	2.18	0.07
	64	0	----	0	----	0	----	2.43	0.07
	65	0	----	0	----	0	----	2.32	0.07

Table 2. Grade 1 estimated slope parameters.

	Item	gamma 1	SE 1	gamma 2	SE 2	gamma 3	SE 3	gamma 4	SE 4	gamma 5	SE 5
Listening	1	3.21	0.09								
	2	1.96	0.06	-0.32	0.07						
	3	-2.65	0.07								
	4	2.05	0.06								
	5	1.19	0.04								
	6	1.07	0.04								
	7	2.69	0.07								
	8	0.81	0.04								
	9	0.44	0.03								
	10	0.84	0.04								
	11	1.35	0.04								
	12	1.06	0.04								
	13	1.18	0.04								
	14	1.07	0.04								
	15	0.49	0.03								
	16	0.41	0.03								
	17	1.28	0.04	0.97	0.05						
	18	1.37	0.05	0.49	0.07	0.51	0.04				
Reading	19	6.14	0.35								
	20	6.99	0.47								
	21	6.9	0.44								
	22	4.06	0.14								
	23	3.55	0.11								
	24	2.51	0.07								
	25	2.68	0.08								
	26	4.82	0.2								
	27	4.06	0.15								
	28	2.16	0.07								
	29	3.1	0.11								
	30	1.97	0.07								
	31	2.2	0.09								
	32	1.62	0.04								
	33	1.05	0.04								
	34	1.02	0.04								
	35	0.79	0.04								
	36	3.18	0.1								
	37	2.31	0.07								
	38	1.89	0.06								
	39	-0.25	0.05								
	40	1.52	0.07								
	41	2.25	0.07								
	42	1.05	0.04								

Writing	43	1.67	0.05								
	44	2.66	0.07								
	45	4.02	0.13								
	46	1.34	0.04								
	47	4.03	0.14								
	48	3.58	0.1								
	49	1.63	0.05								
	50	-0.56	0.04								
	51	0.53	0.04								
	52	0.03	0.04								
	53	-0.83	0.05								
	54	-1.74	0.05								
	55	1.39	0.05								
	56	-0.83	0.04								
Speaking	57	1.47	0.04	0.8	0.05						
	58	1.1	0.03	0.07	0.04						
	59	1.47	0.05	2.39	0.09	0.06	0.03				
	60	2.15	0.07	2.31	0.1	0.15	0.04				
	61	1.8	0.06	2.58	0.1	0.01	0.03				
	62	1.18	0.04	2.03	0.07	0.11	0.03				
	63	0.75	0.06	6.64	0.2	0.74	0.05	-0.02	0.04	-0.03	0.03
	64	1.41	0.08	7.9	0.24	0.63	0.06	0.09	0.04	0.15	0.03
	65	1.14	0.07	7.7	0.24	0.33	0.06	-0.02	0.04	0.04	0.03

Table 3. Grade 1 estimated intercept parameters.

Appendix 2

Table 2.5 Scoring Rubric for the Writing Performance Domain Component

Rating	Vocabulary	Sentence Fluency	Grammar	Mechanics	Organization and Development
4	Descriptive and vivid vocabulary is specific and enhances meaning.	The writer uses a variety of sentence structures, including compound and/or complex sentences, correctly and effectively.	Verb tenses and subject/verb agreement are varied, appropriate, and correct. Any errors that are present do not impede meaning.	There are few or no errors in punctuation, capitalization, and spelling. The writer uses indentation and alignment for paragraph format.	There is a clear main idea with details that are relevant, specific, and appropriate to the prompt. Supporting ideas are well organized with a clear beginning, middle, and end. Transitions are used correctly.
3	Vocabulary is adequate but not expressive; words are more general than specific. A few usage errors may be present but do not impede communication.	Sentences have some variety and may attempt compound structures. A few errors may be present, but do not impede communication.	Writing may contain verb tense and subject/verb agreement errors. Errors do not interfere with meaning.	Some errors in punctuation, capitalization, and spelling are present.	The paper is focused on a single idea or event related to the prompt, but may provide only simple supporting details as well as some extraneous details. Writing has a recognizable flow from start to finish, but there are few or repetitive transitional signals.
2	Vocabulary is limited with usage errors that may impede communication.	Sentences consist of basic structural patterns and may contain errors that may impede meaning. Sentence construction may be awkward.	Writing contains frequent errors in verb tenses and subject/verb agreement. Errors might interfere with meaning.	Frequent errors in punctuation, capitalization, and spelling.	There is a main idea or focus, but details are unrelated.
1	Vocabulary is limited to isolated words and/or phrases.	There are no complete sentences.	Writing contains consistent errors in verb tense that impede understanding.	Consistent errors in punctuation, capitalization, and spelling.	Writing only consists of repetitive simple sentences, fragments, or isolated words.
0	There is no response or the response is mostly in a language other than English.	There is no response or the response is mostly in a language other than English.	There is no response or the response is mostly in a language other than English.	There is no response or the response is mostly in a language other than English.	There is no response or the response is mostly in a language other than English.

Table 1. Scoring Rubric for the Writing Performance Domain Component.

Scoring Guidelines for Speaking

Scoring Rubric #1 for Items 1-4

Score the student's response (compared to native English-speaking peers) to items 1-4 using the rubric scale below. Record the score for **EACH** item on the students' answer sheets during testing. Score as follows:

2 = Any appropriate response using at least **ONE** complete and grammatically correct sentence

1 = Any appropriate response in a phrase/any appropriate response using one word/any appropriate response with grammatical errors

0 = No response or inappropriate response to prompt to response in native language

Scoring Rubric #2 for item # 5

Score the students' responses to item #5 using the rubric scale below and record the score for the appropriate item on the students' answer sheets during testing. Score as follows:

5 = Speech approximate to a native English speaker: provides an elaborate story or detailed description of a prompt with virtually no errors

4 = Provides a story or description of a prompt using three or more complex sentences with some detail; a few errors may be present that do not affect meaning

3 = Provides a basic story or description of a prompt using two to three sentences; some errors may be present that do not affect meaning

2 = Uses one to two simple sentences related to a prompt; errors may be present that might affect meaning

1 = Names objects, people, or actions using only single words and phrases

0 = No response or disconnected oral remarks or response in native language

Retrieved from http://www.title3.greenbush.us/SpeakingRubric.pdf on 10/1/2014.