ISSUES AND CHALLENGES IN CURRENT GENERALIZABILITY THEORY APPLICATIONS IN RATED MEASUREMENT

BY

CHIH-KAI LIN

DISSERTATION

Submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy in Educational Psychology
in the Graduate College of the
University of Illinois at Urbana-Champaign, 2014

Urbana, Illinois

Doctoral Committee:

Associate Professor Jinming Zhang, Chair, Co-Director of Research
Professor Fred Davidson, Co-Director of Research
Professor Katherine Ryan
Professor Carolyn J. Anderson

ABSTRACT

The current dissertation looks into issues and challenges regarding the use of generalizability theory or G theory (Brennan, 2001; Cronbach, Gleser, Nanda, & Rajaratnam, 1972; Shavelson & Webb, 1991) in rated measurement given by human raters. Contexts in which such measurement prevails include, but are not limited to, performance-based assessments, standard settings, and content validation studies. Inherent in expert rated measurement are potential systematic and random variations that can contribute to measurement errors, and thereby affect measurement reliability. Examples of systematic variability (i.e., *facets* in G-theory terminology) are differences in rater severity/leniency, variations in rater interpretations of scoring criteria, and interactions of these facets with the objects of measurement (i.e., the subjects on which the intended construct is measured), whereas random variability reflects unexpected fluctuations in the rating process. Given that the utility of any rated measurement is contingent upon its reliability, analytical tools for disentangling variability in the objects of measurement from variations associated with measurement facets and associated with random errors are necessary. To this end, G theory provides a powerful analytical framework that allows investigators to tease out true differences among the objects of measurement and to assess the relative magnitude of construct-irrelevant variability.

This dissertation follows a multi-paper approach and includes six chapters, including an introduction, four individual papers pertaining to theoretical and applied investigations of G theory in rated measurement, and a conclusion. The introduction (Chapter 1) sketches an overarching theme that situates the separate papers in a thematic unity and also provides a brief summary of each paper. Next, the first paper (Chapter 2) reports on findings from comparing two

analytical methods, under the G-theory framework, which are designed to analyze sparse rated data commonly observed in performance-based assessments. The *rater* method identifies blocks of fully crossed sub-datasets and then estimates variance components based on a weighted average across these sub-datasets, while the *rating* method forces a sparse dataset to be a fully crossed one by conceptualizing ratings as a random facet and then estimates variance components by the usual crossed-design procedures. This paper aims to compare the estimation precision of the two methods via a Monte Carlo simulation study and an empirical study. Results show that when all raters are expected to be homogeneous in their score variability, either method has good estimates of variance components. However, when some raters exhibit more variability in their ratings than others, the rater method yields more precise estimates than the rating method.

The second paper (Chapter 3) is carried out in the context of examining correspondence between English language proficiency (ELP) standards and academic content standards in the US K-12 setting. Such correspondence studies provide information about the extent to which English language learners are expected to encounter academic language use closely associated with academic disciplines, such as mathematics. This paper describes one approach to conducting ELP standards-to-standards correspondence research based on reviewer judgments, and it also touches on reviewer consistency in judging the cognitive complexity of the target standards. Results suggest that there seems to be a relationship between reviewer consistency in their judgments and the level of specificity in the target standards. As an extension of the second paper, the third paper (Chapter 4) seeks to advance new applications of G theory in correspondence research and to examine reviewer reliability in relation to the numbers of raters. Ratings of the cognitive complexity germane to language performance indicators were collected

from 28 correspondence studies with over 700 trained reviewers, consisting of content-area experts and English as a second language (ESL) specialists. Under the G-theory framework, reviewer reliability and standard errors of measurement in their ratings are evaluated with respect to the numbers of reviewers. Results show that depending on the particular grades and subject areas, 3-6 reviewers are needed to achieve an acceptable level of reliability and to control for a reasonable amount of measurement errors in their ratings.

The fourth paper (Chapter 5) attempts to advance the discussion of nonadditivity in the context of G-theory applications in rated measurement. Nonadditivity occurs when some or all of the main and interaction effects, pertaining to the objects of measurement and measurement facet(s), are significantly correlated. As such, the paper analytically and empirically illustrates the distinction between additive and nonadditive one-facet G-theory models. In addition, the paper aims to explore existing statistical procedures of detecting nonadditivity in data. Tukey's single-degree-freedom test for nonadditivity is evaluated in terms of Type I error and statistical power. Results show that the test is satisfactory in controlling for occurrences of erroneously identifying nonadditivity (Type I error) and that the test is successful in identifying one type of nonadditive interaction (power).

As will become clear in the dissertation, the first and fourth papers are motivated by methodological challenges in advancing G-theory applications in the field of educational measurement, while the second and third papers are motivated by validity issues in assessing the content knowledge of young English language learners in the field of language testing. Finally, the conclusion (Chapter 6) functions as a discussion of some unsolved issues in G-theory applications and ideas for future research. First, issues regarding the use of many-facet Rasch measurement to complement G-theory analysis are discussed. Second, given that a performance

test usually involves examinee responses being rated on a discrete ordinal scale, the consideration of the discrete ordinal nature in measurement variables under the G-theory framework is an unsolved area of research. Finally, nonadditivity in multi-faceted G-theory models is also an area that deserves more research efforts because most performance tests would entail more than one measurement facet, such as those associated with raters and tasks.

ACKNOWLEDGMENTS

I would like to express my profound gratitude to my dissertation committee—Jinming Zhang, Fred Davidson, Katherine Ryan, and Carolyn Anderson—for their invaluable guidance and support. The current dissertation would not have been possible without the constructive discussions with my committee members. Fred Davidson has been a great and caring mentor at a professional as well as a personal level. I also owe a deep debt of gratitude to Jinming Zhang for our fruitful discussions that led to research ideas and for our casual conversations that cheered me up. The Wisconsin Center for Education Research (WCER) and the English Placement Test (EPT) program at the University of Illinois at Urbana-Champaign are also acknowledged for providing datasets necessary to conduct the dissertation. Rika Kinoshita deserves special thanks. She has supported me through the course of my graduate-school career and has shown great patience during times when this journey seemed endless. Finally, I would like to thank my family in Taiwan for believing in me.

TABLE OF CONTENTS

CHAPTER 1

INTRODUCTION

Expert rated assessments of actual test performances are common in a plethora of contexts, such as academic departments at universities that rely on placement tests to assess incoming students, regional and national governments that administer achievement tests to measure student growth, large-scale testing programs that offer academic and work-place qualifications, and researchers who need performance tests for the purpose of their research. The advent of performance-based tests is partly driven by validity concerns regarding the extent to which assessment tasks resemble real-world tasks and the degree to which test performances can be safely generalized to non-test contexts, which are in accord with the modern paradigm of test validation (Kane, 2006; Messick, 1989).

Given the emphasis on performance tests, rater-mediated measurement has become typical in many assessment systems; in addition, many testing programs continue to rely on a time-honored scoring paradigm: expert raters with rigorous training and calibration. However, scoring test performances by human raters comes with a set of stress factors. For example, even in a well-designed rating system, certain practical realities might mitigate the effectiveness of rater training, such as time pressure due to a short turnaround timeline for scoring. Furthermore, some raters may resign or be ill, forcing test administrators to use a smaller pool of trained raters or to turn to a wider pool of former raters, some of whom have not been fully or recently re-calibrated. All of these factors result in score fluctuations for reasons other than the intended construct being measured.

Other areas in which rater-mediated measurement prevails include, but are not limited to, standard-settings studies and content-validation studies. The purpose of standard settings is to establish cut-off proficiency levels or standards for achievement, licensure, and certification tests. Generally, panels of expert raters participate in this judgmental process (Cizek, Bunch, & Koons, 2004; Kane; 1994; Nichols, Twing, Mueller, & O'Malley, 2010). In content-validation studies, the alignment between a test and its specification, from which the test is derived, serves as a necessary piece of validity evidence in relation to inferences drawn from test scores. Content alignment is usually investigated based on judgments from content experts (Davidson, 2012; Davidson & Fulcher, 2012; Martone & Sireci, 2009; Sireci, 1998). The aforementioned set of stress factors for human raters in performance-based assessments may also be present for those in these other types of rater-mediated measurement. This leads to the central theme of the current dissertation: the use of humans as judges given existing realities in rated measurement.

The current dissertation adopts a multi-paper approach by which a number of papers with different topics are connected by the thematic unity. In particular, the first paper pertains to the existing reality of sparse rated data as a result of operational design constraints on scoring performance data by human raters. The second and third papers touch on the existing reality of recruiting expert reviewers in standard-based judgmental research. The fourth paper deals with the existing reality of potential person-by-facet interactions and their impact on the analysis of rated measurement. In addition to its thematic unity, the dissertation operates under a unified generalizability theory (G theory) framework in that the use of G theory is uniquely motivated by research questions raised in the papers. G theory has been widely used in capturing proportions of total score variance explicable by various sources of measurement variability and in examining score reliability with respect to different measurement designs. More details on G

theory as a analytical framework and on G-theory applications in education are provided in Chapters 2-5 of the dissertation.

The first and fourth papers are related in the following way: recognizing and coping with patterns of rated datasets as a given, whereas the second and third papers work with a quite different type of measurement: conducting judgmental studies of the association between educational standards. As such, the dissertation is related to the diverse contexts presented in the individual papers, from which real-world implications are derived. Furthermore, all four papers touch on reliability in rater-mediated measurement, although the topic of reliability is not necessarily the major focus of each paper. Reliability here is not interpreted as the internal consistency of items/tasks in a typical item analysis—the degree to which items/tasks correlate with each other and jointly measure a defined construct (Allen & Yen, 2001), nor is it conceptualized as a layperson interpretation of trustworthiness—the extent to which the measurement is accurate (Ennis, 1999). Rather, reliability in rater-mediated measurement is about the extent to which raters are consistent in giving scores across the objects of measurement (e.g., examinees and performance descriptors) according to a rating rubric (Stemler & Tsai, 2008). Rater-mediated measurement is a product of raters' understanding of the intended construct being measured, their interpretations of the rating rubrics, and their use of the rubrics in making their judgments. High inter-rater reliability is desirable so that raters can be considered interchangeable; that is, a score awarded is not contingent upon the specific raters that are assigned to make the judgment. Next, a brief summary of each paper is provided.

The first paper (Chapter 2), entitled “Estimating variance components from sparse data in rated language tests: A simulation study,” is motivated by the pervasiveness of sparse data in language performance tests due to real-world design constraints on scoring performance data by

human raters. It compares two analytical methods of estimating variance components from sparse data in performance-based language assessments under the G-theory framework. Investigating the precision of estimated variance components is of great importance given that these estimates function as building blocks for computing score reliability, on which valid inferences drawn from scores of any performance-based language assessments depend. First, the rater method identifies blocks of fully crossed sub-datasets and then estimates variance components based on a weighted average across these sub-datasets (e.g., Xi, 2007). Second, the rating method forces a sparse dataset to be a fully-crossed one by conceptualizing ratings as a random facet and then estimates variance components by the usual crossed-design procedures (e.g., Bachman, Lynch, & Mason, 1995; Huang, 2012; Lee & Kantor, 2005). This paper aims to compare the estimation precision of the two methods via a simulation study. Results show that when all raters are expected to be homogeneous in their score variability, either method has good estimates of variance components. However, when some raters exhibit more variability in their ratings than others, the rater method yields more precise estimates than the rating method. Implications for methodological approaches to handling sparse data are discussed. Finally, the paper demonstrates applications of the two methods in analyzing an operational sparse dataset from a university English writing placement test.

The second paper (Chapter 3), entitled "English language proficiency (ELP) standards-to-standards correspondence research: Reviewer consistency," highlights the importance of ELP correspondence research by describing one approach to investigating the expected academic language loads residing in academic content standards (Cook, 2006, 2007). By drawing on a study about correspondence between the WIDA English Language Proficiency Standards (World-class Instructional Design and Assessment, 2007) and the Common Core State Standards

for Mathematics and English Language Arts (Common Core State Standards Initiative, 2010a, 2010b), this paper also touches on reviewer consistency in interpreting the standards; more importantly, it offers recommendations for planning ELP correspondence studies based on empirical quantitative and qualitative results. First, investigators should strive to identify reviewers who have sufficient training and experience working with the target standards so that reliable judgments among panels of reviewers can be better assured. Second, the balance between recruiting content experts and recruiting English as a second language (ESL) specialists could help boost reviewers' awareness of academic language use inherent in subject areas. Third, give the target standards, ELP standards-to-standards correspondence studies should be conducted at the highest possible level of specificity of the standards so that variability in reviewer interpretations of the standards could be reduced.

The third paper (Chapter 4), entitled "Dependability of reviewer judgments about language performance indicators: How many reviewers?" concerns reviewer reliability in the study of correspondence between ELP standards and academic content standards in the US K-12 setting. ELP standards-to-standards correspondence is an area of emerging significance in that it broadens the notion of content alignment to better serve the English language learner (ELL) populations in a standard-based assessment system (Bailey & Butler, 2004; Bailey & Huang, 2011; Bailey, Butler, & Sato, 2007). In addition, correspondence research can provide information about the extent to which ELLs are expected to encounter academic language use that facilitates their content learning, such as in mathematics and science. Standards-to-standards correspondence thus contributes to standard-based validity evidence regarding ELLs' achievement levels in content areas. This paper examines the reliability of reviewer judgments about language performance indicators associated with academic disciplines in standards-to-

standards correspondence studies. Ratings of cognitive complexity germane to the language performance indicators were collected from 28 correspondence studies with over 700 content experts and ESL specialists as reviewers. Under the G-theory framework, reviewer reliability and standard errors of measurement in their ratings are evaluated with respect to the numbers of reviewers. Results show that depending on the particular grades and subject areas, 3-6 reviewers are needed to achieve an acceptable level of reliability and to control for a reasonable amount of measurement errors in standards-to-standards correspondence studies.

The fourth paper (Chapter 5), entitled "Evaluating Tukey's single-degree-freedom test for detecting nonadditivity in rated measurement," is derived from an unpublished manuscript by Zhang and Lin (2013). In their paper, the authors discussed differences and implications for additive and nonadditive G-theory models. For instance, a fully-crossed ($p \times r$) design corresponds to the following linear model: $X_{pr} = \mu + \alpha_p + \beta_r + \varepsilon_{pr,e}$, where the score ($X_{pr}$) of person p given by rater r is the sum of an overall mean (μ) and the three random effects pertaining to persons, raters and errors. The current G-theory framework assumes that "all effects in the model are uncorrelated" (Brennan, 2001, p. 23), which is equivalent to the assumption of additivity. Nevertheless, the additive assumption may not always hold in operational settings, particularly in the presence of person-by-facet interaction effects. When such interaction effects exist, the person effect is likely to be correlated with the interaction, resulting in nonadditivity. The authors, in their unpublished manuscript, have shown that when additive models are inadvertently used to analyze nonadditive data, the variance component of persons can be adversely underestimated, leading to possible negative variance component estimates, which are at odds with the notion of variance components. As a follow-up to this research, identification of nonadditivity from data is of great importance in that the selection of appropriate G-theory

models, be it additive or nonadditive, needs to be informed by statistical tests for nonadditivity. To this end, Tukey's single-degree-freedom test for nonadditivity (Tukey, 1949) is one candidate that is appropriate for the typical G-theory designs in which a single observation is made per element within a facet. The current paper aims to evaluate Tukey's test in terms of Type I error and power via a Monte Carlo simulation study. Results indicate that Tukey's test is able to control for reasonable Type I error rates and to achieve satisfactory statistical power. More importantly, the paper demonstrates an application of Tukey's test in a judgmental study of educational standards and provides relevant computer program syntax for performing Tukey's test.

The following four chapters present the four individual papers, which are headed by their respective titles. For readers' convenience, each of the four papers is written in a self-contained fashion such that a proportion from one paper may be reiterated in other paper(s) due to the inter-relatedness among some parts of the papers. Also note that figures, tables, and equations are numbered by means of a decimal convention: the first digit refers to the chapter number and the digit after the decimal point is the figure/table/equation number within each chapter. Following the four papers, the dissertation concludes with a discussion of some unsolved issues in G-theory applications and ideas for future studies (Chapter 6).

CHAPTER 2

ESTIMATING VARIANCE COMPONENTS FROM SPARSE DATA IN RATED LANGUAGE TESTS: A SIMULATION STUDY

Performance-based language testing serves as an alternative to traditional multiple-choice item formats. It offers a more direct measure of a person's proficiency in language domains of speaking and writing. The advent of performance-based language testing is partly motivated by validity concerns regarding the extent to which test performances can be generalized to target language use in non-test settings (Bachman & Palmer, 1996; Chapelle, Enright, & Jamieson, 2008). Moreover, valid inferences about examinees are contingent upon score reliability. Hence, analytical approaches to examining score reliability are needed to determine the utility of any language performance tests.

Various sources of systematic variability can contribute to measurement errors in performance-based language testing, and thereby affect score reliability. These sources include, but are not limited to, rater severity, task difficulty, topic familiarity, scoring rubrics, and testing conditions (Schoonen, 2005). To account for these variations and others not identified here, useful candidates for measurement models need to be able to capture potential sources of systematic variability inherent in performance-based language assessments. To this end, generalizability theory or G theory (Brennan, 2001; Cronbach, Gleser, Nanda, & Rajaratnam, 1972; Shavelson & Webb, 1991) is a powerful analytical tool that has been successful in investigating score reliability with respect to multiple sources of systematic variations in language performance tests (e.g., Akiyama, 2001; Bachman, Lynch, & Mason, 1995; Brown & Ahn, 2011; Elorbany & Huang, 2012; Gebril, 2009; Huang, 2012; Huang & Foote, 2010; Kim,

2009; Lee, 2006; Lee & Kantor, 2005, 2007; Lynch & McNamara, 1998; Xi, 2007). Nevertheless, additional complications of G-theory applications need to be addressed when it comes to sparse datasets in a rated test. Particularly in the context of language testing where the use of rated performance data has become typical in many testing programs, it is fairly common to have sparse data in practice due to real-world design constraints on scoring language performance data by human raters. Sparse data will be discussed later in more details following a brief introduction to G theory.

2.1 A Brief Introduction to Generalizability Theory

In classical test theory (Lord & Novick, 1968), an examinee's observed score is conceptualized as a composite of a true score and a single error term. The true score is the hypothetical expected score of the examinee being measured repeatedly an infinite number of times, and the single error term encompasses potential systematic and random errors. G theory (Brennan, 2001; Cronbach et al., 1972; Shavelson & Webb, 1991) expands the classical test theory by re-conceptualizing the undifferentiated error component into different sources of systematic variability (i.e., facets) and into random errors. This decomposition provides information about how much variation is explicable by each facet. Estimated variance components pertaining to the objects of measurement (i.e., persons or examinees), facet(s), and errors are then used as building blocks to examine reliability-like coefficients and standard errors of measurement (SEMs).

Conceptually, G theory distinguishes between generalizability studies (G studies) and decision studies (D studies). In G studies, investigators first specify measurement facet(s) that may be of interest and/or of great influence on observed measurements. Variance components of the main and interaction effects pertaining to the objects of measurement and the facet(s) are

then estimated. These estimates translate to fractions of total score variance attributable to each facet. In D studies, G-study results (i.e., variance component estimates) are to be generalized from a particular measurement procedure, usually the data collection procedure at hand, to other measurement procedures. Specifically, D studies can provide information as to how score reliability and SEMs change when the number of ratings/raters increases or decreases. In other words, D studies are for decision making in terms of effective measurement designs (Brennan, 2000).

2.2 Analyzing Sparse Data under G-Theory Framework

Many language assessments elicit writing and speaking performances that are assessed by human raters. The prevalence of these rated measurements necessitates the use of appropriate analytical tools, such as G theory which can parse out (and account for) the variability in the raters' scores. Despite its popularity in examining score reliability of a rated language test, ideal applications of G theory require fully-crossed measurement designs, meaning that persons are crossed with facet(s) involved in the measurement. In such ideal designs, the main effects of persons and individual facets can be estimated separately; otherwise, some effects may be confounded with others in the analysis of sparse rated data.

For example, a fully-crossed $(p \times r)$ design in a writing test requires that each written response or person (p) to be scored by all raters (r). However, fully-crossed designs are oftentimes not feasible in operational settings due to logistical concerns and/or structural difficulties. As pointed out by Lee (2006) that when fully-crossed designs are sought, "each rater is required to rate an unrealistically large number of performance samples on multiple tasks for examinees in a single rating session" (p. 138). Simply put, many language test administrations cannot afford to have each rater score all examinee responses. Take a university English writing

placement test for example. Turnaround time for placement results may need to be rather fast so that placement decisions can be made promptly in order for students to register for appropriate courses. Given that, it can be more cost-effective to assign different batches of responses to groups of raters; as a result, sparse rated data are inevitable in practice. In typical double-rating procedures, one simple example of sparse data is from a nested $(r:p)$ design, in which raters are nested within examinees such that different groups of examinees are scored by different pairs of raters. Other more complex sparse data entail cross-pairing of raters and overlapping of raters for different groups of examinees. The current study focuses on double-rating procedures with both simple and complex sparse datasets.

In ideal crossed designs, variance components associated with the main effects of persons and individual facets can be estimated separately; however, this is not the case with sparse data. For instance, in a nested $(r:p)$ design, the rater facet is confounded with errors and cannot be estimated independently of the error component. In pursuit of separate variance-component estimation from sparse data, two methods based on Analysis of Variance (ANOVA) procedures have been applied under the G-theory framework in performance-based assessments. These variance component estimates are then used to compute score reliability. First, *raters* are treated as a random facet (e.g., Xi, 2007); henceforth referred to as the rater method. Second, *ratings* are treated as a random facet (e.g., Bachman, Lynch, & Mason, 1995; Huang, 2012; Lee & Kantor, 2005); henceforth referred to as the rating method. The two methods differ not only in the specification of the random facet but also in the estimation procedures of variance components. Figure 2.1 gives a visual representation of how the two methods break down a hypothetical sparse dataset, where each response from sixty persons/examinees (P1-P60) is double-rated among a panel of four raters (R1-R4), into crossed dataset(s). As such, the rater method first

identifies all possible blocks of fully crossed sub-datasets and estimates the variance components in each block. These variance component estimates are then averaged across the sub-datasets by giving weights according to the number of examinees in each block (Chiu, 2001; Chiu & Wolfe, 2002). The rating method forces a sparse dataset to be a fully crossed one by conceptualizing individual ratings, irrespective from which raters, as a random facet. The variance components are then estimated via the usual ANOVA procedures for any fully-crossed designs.

Clearly, the rating method is computationally less complex but achieves its simplicity at the expense of rater information by assuming that score variability is similar across all raters, which may not always be the case when a mixture of novice and experienced raters participate in scoring. The rater method retains rater information by assigning different weights to groups of raters; nevertheless, it requires higher computational sophistication, particularly when the number of all possible crossed blocks is large. The two methods have been successful in analyzing sparse data from performance-based language tests. For instance, the rating method has been applied in a university Spanish placement test (Bachman, Lynch, & Mason, 1995), in an achievement test for secondary-school ESL students (Huang, 2012), and in an English language test for immigration purposes (Lynch & McNamara, 1998). The rater method has been applied in a large-scale English language proficiency test (Xi, 2007). In spite of the usefulness of the two methods, little research has directly compared the different methods of handling sparse datasets in terms of estimation precision under the G-theory framework. The precision of estimated variance component is of great importance in that these estimates are directly involved in the computation of score reliability. The current study aims to fill this gap by investigating the precision of the rating and rater methods in estimating variance components.

2.3 Research Questions

Researchers/practitioners who work with sparse rated data have the rating and rater methods at their disposal in performing G-theory analyses. Given the two methods, the issue here does not merely rest on computational sophistication or the amount of rater information; rather, the fundamental question is whether the methodological approaches can achieve precise estimation of variance components given that these estimates are used to compute score reliability. The current study compares the estimation precision of variance components based on the rating and rater methods, and it also demonstrates applications of the two methods in analyzing rated performance data from a university English writing placement test. Specifically, the following four research questions are addressed:

1. When raters exhibit similar variability in their scoring, does one method yield more precise estimates of variance components than the other based on sparse data?
2. When raters have varying degrees of score variability, does one method yield more precise estimates of variance components than the other based on sparse data?
3. Under what rating condition(s) is one method recommended over the other?
4. How do different estimation procedures from the two methods impact variance component estimates, reliability estimates, and SEMs from a university English writing placement test?

This study is method-oriented and yet with real-world implications for language testers who work with rated performance data in that it offers methodological recommendations for analyzing sparse rated data from language performance tests.

2.4 Method

The current paper includes a simulation study and an empirical study. The goal of the simulation study was to investigate estimation precision (i.e., research questions 1, 2, and 3), whereas the goal of the empirical study was to apply the rating and rater methods in operational data analysis to address research question 4. Although an empirical study by Lee and Kantor (2005) has shown that similar estimated variance components were observed either when a fully-crossed design was employed or when ratings were treated as a random facet, baseline comparisons with true variance parameters are not possible in empirical research because the true parameters are unknown. The current paper builds on this line of empirical research and expands the scope via a Monte Carlo simulation study, in which the true variance parameters are predetermined, to compare the estimation precision of the two methods with respect to analyzing sparse data.

For each simulated and operational dataset, the measurement design followed a fully-crossed ($p \times r'$) one-facet random effect model under the rating method, where p is the objects of measurement and r' refers to the random facet of ratings. Under the rater method, a sparse dataset was decomposed into all possible blocks of fully-crossed sub-datasets, and each sub-dataset followed a ($p \times r$) one-facet random effect model, where r refers to the random facet of raters.

2.4.1 Simulation study. Another advantage of using simulation procedures to investigate estimation precision is that, instead of operating under a single operational setting in which an empirical study is usually carried out, investigators can purposefully choose simulated conditions that mirror multiple realistic settings, providing useful implications for a wider audience of practitioners who work in diverse contexts. Three target sample sizes (n_p): 50, 100 and 200, three numbers of raters: 4, 8 and 16, and three scenarios of rater score variability were chosen;

hence, a total of 27 conditions were considered in the simulation study. The three rater scenarios were: (a) all raters exhibit similar variability in their scoring, (b) a small minority of raters has greater score variability than the rest, and (c) a large majority of raters exhibits greater variability in their scoring than the rest. These conditions were intended to reflect realistic settings in terms of sample sizes, numbers of raters, and rater compositions. For instance, the sample size of 50 corresponded to a typical number of examinees taking a language placement test in a single test session for a second/foreign language education program, while the scenarios (b) and (c) mirrored rating designs in which a mixture of novice and experienced raters participate in a single test session. For clarity, a step-wise overview of the simulation procedures is sketched here. First, a sparse dataset was generated according to a particular simulated condition. Second, variance component estimates were obtained based on the rating and rater methods, respectively. Third, these estimates were evaluated against their corresponding true variance parameters. The above three steps were repeated 1,000 times for each condition so that general trends of estimation precision for the rating and rater methods can be compared. Next, technical details regarding each simulation step are provided.

Data associated with the scenario (a) were generated according to a one-facet random effect model:

$$X_{pr} = \mu + \alpha_p + \beta_r + \varepsilon_{pr,e}. \quad (2.1)$$

For example, the writing score (X_{pr}) of person p given by rater r is the sum of an overall mean (μ) and the three random components pertaining to persons, raters and errors. These three random components were generated independently from three normal distributions, where the person effect (α_p), the rater effect (β_r), and the error component ($\varepsilon_{pr,e}$) follow a normal distribution with a mean of zero and variance of σ_p^2, σ_r^2, and σ_e^2, respectively. A single value

generated randomly from the person distribution can be thought of as the relative standing of that person to a person with average writing proficiency. In a similar vein, a random value from the rater distribution is the standing of that rater relative to a rater with average rating severity/lenience. A random value from the error distribution reflects score fluctuations by chance. Examinee scores were simulated to be scored on a five-point scale (from 1 to 5), and therefore the overall mean was set at 3. Take one of the simulated conditions—n_p = 200, n_r = 16, and rater scenario (a)—as an example. A 200-by-16 dataset generated by Model (2.1) under this condition is equivalent to writing scores of a random sample of 200 examinees given by a random sample of 16 raters who have similar score variability. It is this randomness involved in data generation that allows the current simulation study to replicate each simulated condition for a large number of times in order to gauge the estimation precision of the rating and rater methods.

True parameters of the variance components—σ_p^2 =.4709, σ_r^2 =.0095, and σ_e^2 =.2223—were selected according to variance component estimates from previous empirical writing studies in which fully-crossed designs were employed (Elorbany & Huang, 2012; Gebril, 2009; Lee & Kantor, 2005, 2007). In particular, estimated variance components from these studies were first adjusted for their scale differences and then averaged across the studies. It should be noted that in a simulation study, true parameters are selected from values that seem reasonable based on previous research (Mooney, 1997). Some G-theory simulation studies adopted values from a single empirical study (e.g., Nugent, 2009) while others heuristically used values from the standard normal distribution (e.g., Tong & Brennan, 2007). The current study attempts to arrive at reasonable true parameters by looking at multiple empirical studies with fully-crossed data and by taking the averages across these studies. These true variance components translate to 67%, 1%, and 32% of total score variance attributable to persons, raters, and errors, respectively. The

justification for the selected true parameters is further supported by the fact that their corresponding proportions of variance components fall within the ranges—63-81%, 1-3%, and 16-36% for σ_p^2, σ_r^2, and σ_e^2, respectively—reported in a meta-analysis of generalizability studies on writing tests (In'nami & Koizumi, 2013).

For the scenarios (b) and (c), data generation also followed Model (2.1), except that the rating variability for novice raters was modeled to be 2 times larger than that for experienced raters. The idea that larger score variability is associated with novice raters was taken from empirical observations that inexperienced raters appeared to be less consistent in their scoring than experienced raters (Weigle, 1998, 1999). Two raters were designated as novice raters across all simulated conditions in the scenario (b), while two raters were experienced raters across all simulated conditions in the scenario (c). Thus, in the scenario (b), novice raters constituted 50%, 25%, and 12.5% of the raters for n_r = 4, 8, and 16, respectively. In a similar vein, experienced raters accounted for 50%, 25%, and 12.5% of the raters for n_r = 4, 8, and 16 in the scenario (c). Given this additional complication, details about true parameters for the overall rater variance components ($\sigma_r^{2\prime}$s) associated with the scenarios (b) and (c) are discussed more fully in section 2.5.1.

In generating sparse datasets, the levels of sparseness were directly linked to the numbers of raters (n_r) because one constraint was imposed such that all raters had the same amount of scoring load. Due to the equal scoring load for each rater in generating the data, the sparse datasets simulated in this study included both simple nested $(r:p)$ design and other more complex designs where one rater is not always paired with another specific rater. An illustration of sparse-data generation is provided here. In the case where n_p=100 and n_r=8, a crossed 100-by-8 dataset with complete data was first generated. The first examinee was randomly assigned

to two raters, and therefore the data for this examinee associated with the other six raters were removed. The next examinee was randomly assigned to two raters, and so on until the constraint of equal scoring load for each rater was met. This resulted in a sparse level of 75%, leading to four 25-by-2 crossed sub-datasets under the rater method and one 100-by-2 crossed dataset under the rating method. In sum, the three numbers of raters (i.e., 4, 8, and 16) corresponded to sparseness levels of 50%, 75%, and 87.5%, respectively.

The estimated variance components produced by the rater and rating methods were evaluated against the true variance parameters with respect to average bias and root mean square error (RMSE) over 1,000 replications of sparse datasets for each condition. For a particular effect f, the average bias and RMSE of its estimated variance component ($\hat{\sigma}_f^2$) were obtained by

$$\text{average bias} = \frac{1}{1000}\sum_{h=1}^{1000}(\hat{\sigma}_{f(h)}^2 - \sigma_f^2) \quad \text{, and}$$

$$\text{RMSE} = \sqrt{\frac{1}{1000}\sum_{h=1}^{1000}\left(\hat{\sigma}_{f(h)}^2 - \sigma_f^2\right)^2} \quad ,$$

where $f = p$ (persons), r (ratings/raters) and e (errors), and h refers to the hth replication. Comparisons between the two methods were possible in that their respective estimation procedures were performed on the same sparse data per condition.

The data generation and variance component estimation were performed in the R statistical software, version 2.15.2. Independent of the current study, estimated variance components were validated with the true parameters by analyzing simulated data with no missing data. The estimated variance components were also verified to be the same as those produced by GENOVA (Crick & Brennan, 1982)—a computer program commonly used in G-study analyses.

2.4.2 Empirical study. One operational dataset, taken from the writing section of the English Placement Test (EPT) administered at the University of Illinois at Urbana-Champaign (UIUC), was analyzed in the empirical study. The purpose of EPT is to place incoming international students at UIUC into appropriate levels of English as a second language (ESL) service courses, which are designed to help international students meet the academic language demands at UIUC. The writing section of EPT consists of an integrated writing task. Test takers are first asked to read an article and listen to a lecture on the same topic but with opposing stances. Next, they are asked to participate in group discussions about the topic. Finally, each test taker writes an argumentative essay by incorporating materials from both the article and the lecture. Each essay is double-scored by trained raters on a three-level scale from 2 to 4, which corresponds to the three levels of undergraduate ESL writing courses offered at UIUC. Given that the three EPT score levels are directly linked to the three ESL placement levels, no intermediate scores are used. Discrepancies between scores are resolved through either consensus discussions or scores from a third rater. EPT raters in the current study were instructors of the ESL writing courses, who were also enrolled in the Teaching English as a Second Language (TESL) master's program at UIUC. Prior to participating in operational scoring, the raters went through a face-to-face training module, in which they worked with the coordinator of ESL writing courses in getting familiar with the EPT scoring procedures, practicing scoring sample essays, and receiving feedback on their scoring.

The EPT data were taken from one of the test sessions in Fall 2012 with 45 test takers and 4 raters. Due to the fact that each essay was not necessarily scored by the same pair of raters, the EPT dataset was sparse rated data. The data were analyzed using both the rater and rating methods. Specifically, the goals of the empirical study were to demonstrate applications of the

different procedures in estimating variance components from the EPT sparse data and to investigate score reliability and SEMs of the EPT writing with respect to the number of ratings/raters. Moreover, absolute interpretations (Brennan, 2001) of reliability and of SEMs were adopted because the scale on which the EPT essays were scored was criterion-based, describing levels of English writing proficiency. Hence, score dependability and SEMs were obtained by

$$\text{phi} - \text{coefficient} = \frac{\hat{\sigma}_p^2}{\hat{\sigma}_p^2 + \frac{\hat{\sigma}_r^2}{n_r'} + \frac{\hat{\sigma}_e^2}{n_r'}} \text{ , and} \tag{2.2}$$

$$\text{SEM} = \sqrt{\frac{\hat{\sigma}_r^2}{n_r'} + \frac{\hat{\sigma}_e^2}{n_r'}} \text{ ,} \tag{2.3}$$

where n_r' refers to the number of ratings/raters in different rating designs.

2.5 Results

2.5.1 Simulation results. Tables 2.1-2.3 are associated with the rater scenario (a), in which all raters exhibit similar variability in their ratings. That is, the raters are matched for their training and/or experience. The three tables present averages, average biases, and RMSEs of estimated variance components for persons ($\hat{\sigma}_p^2$), ratings/raters ($\hat{\sigma}_r^2$), and errors ($\hat{\sigma}_e^2$), respectively. Within each row of n_r, the upper row shows results from the rating method while the lower row represents those from the rater method.

From Tables 2.1-2.3, results show that the average estimates of each variance component produced by the rating and rater methods are almost identical under the nine simulated conditions (3 sample sizes crossed with 3 numbers of raters). The average biases of each estimated variance component based on the two methods do not differ much from each other and are fairly small. In addition, the RMSEs are similar and within reasonable ranges. It is important

to note that average biases and RMSEs both serve as useful indices in gauging estimation precision. The main difference between the two is that biases do not take into account that under- and over-estimations may cancel each other out, whereas RMSE computations are based on squared deviations and therefore reflect deviations from the true parameters in an absolute sense. As expected, when holding the number of raters constant, the RMSEs of each estimated variance component decrease as the number of examinees increases. Take Table 2.1 (i.e., the person effect) for example. When n_r=4, the RMSEs based on the rating and rater methods decrease from .1197 and .1210 to .0594 and .0596, respectively, as n_p increases from 50 to 200. Nonetheless, when the number of examinees is fixed, the RMSEs do not change much as the number of ratings/raters increases.

Tables 2.4-2.6 pertain to the rater scenario (b), in which a small minority of raters has greater score variability than the rest. That is, more experienced raters participate in scoring. Given the varying degrees of score variability across raters, the true parameter $\sigma^2_{r(exp)}$ for experienced raters followed the empirically-driven value at .0095, while the true parameter $\sigma^2_{r(nov)}$ for novice raters was set to be twice of $\sigma^2_{r(exp)}$ at .0190. Due to this setting, true parameters for the overall rater variance components ($\sigma_r^{2\prime}$s) vary as the combination of novice and experienced raters changes, and these parameters cannot be analytically obtained. However, it can be approximated by simulations over a large number of replications. Take n_r=8 for example. To approximate the true parameter σ_r^2 by simulations, a dataset with no missing data was first generated based on the one-facet random Model (2.1), with individual rater effects of the 6 experienced raters following $N(0, \sigma^2_{r(exp)} = .0095)$ and individual rater effects of the 2 novice raters following $N(0, \sigma^2_{r(nov)} = .0190)$. The overall variance component for the rater effect was then estimated from the full dataset. The above process was independently repeated

10,000 times in order to arrive at a stable approximation of the true parameter σ_r^2 by taking the mean across the 10,000 replications. This approximation procedure was performed for each of the three numbers of raters in the scenarios (b) and (c). In the scenario (b), the approximated true parameters for the overall rater variance components (σ_r^2's) were .01425, .01188, and .01069 for n_r=4, 8, and 16, respectively.

As can be observed from Table 2.4 for the person effect, the average variance component estimates are similar based on the two methods. They also converge to the true parameter; as a result, the average biases and RMSEs are small for the person variance component. The same goes with Table 2.6 for the estimated error variance component, although the rating method (i.e., upper row) consistently yields greater magnitude of average biases than the rater method (i.e., lower row). On the other hand, the average estimates for the rating/rater effect differ based on the two methods. $\hat{\sigma}_r^2$'s from the rating method show a larger extent of bias than those from the rater method. Take Table 2.5, where n_r=8 and n_p=200, for instance. The approximated true parameter σ_r^2 is .01188. The rating method underestimates σ_r^2 by .00534 on average, whereas the rater method is short by only .00003. Lastly, although the average biases of $\hat{\sigma}_r^2$ and $\hat{\sigma}_e^2$ based on the rating method are greater in magnitude, it is expected that the degree of bias decreases as the number of ratings increases. For example in Table 2.5, the average biases for n_p=100 change from -.00725 to -.00324 as the number of ratings increases from 4 to 16, and in Table 2.6, the average biases for n_p=50 decrease from .0072 to .0032 as the number of ratings increases from 4 to 16.

Tables 2.7-2.9 are associated with the rater scenario (c), in which a large majority of raters exhibits greater variability in their scoring than the rest. That is, more novice raters participate in scoring. The three tables present averages, average biases, and RMSEs of estimated

variance components for persons, ratings/raters, and errors, respectively. Again, within each row of n_r, the upper row represents results from the rating method while the lower row shows results from the rater method. Following the approximation procedure previously described, the approximated true parameters for the overall rater variance components (σ_r^2's) were .01424, .01663, and .01781 for n_r=4, 8, and 16, respectively.

Generally, when the number of raters is fixed, the RMSEs decrease as the number of examinees increases for each estimated variance component. For the person effect in Table 2.7, results suggest that the average variance component estimates based on the two methods do not differ much from each other and are also close to the true parameters; hence, the average biases and RMSEs are fairly small for the estimated person variance component. The same goes with the estimated error variance component in Table 2.9. However, as can be observed from Table 2.8, the average $\hat{\sigma}_r^2$'s based on the two methods do not conform. Again, the magnitude of average biases based on the rating method is larger than that based on the rater method for the rating/rater effect. For example in Table 2.8, where n_r=16 and n_p=50, the approximated true parameter σ_r^2 is .01781. The average bias from the rating method is -.00337, whereas the average bias from the rater method is -.00043. In addition, although the magnitudes of average biases for $\hat{\sigma}_r^2$ and $\hat{\sigma}_e^2$ based on the rating method is consistently greater than those based on the rater method, the degree of bias decreases as the number of ratings increases.

2.5.2 Empirical results. Sample means, standard deviations, and ranges for the writing scores of EPT are reported in Table 2.10. The descriptive statistics here are based on the final EPT writing scores averaged across pairs of raters. The mean of EPT scores for this sample is 2.51 on the three-level scale of 2-4. Tables 2.11 shows the estimated variance components and their proportions of total variance based on scores from the writing component of EPT. From a

methodological perspective, the results indicate that the rating and rater methods yield similar proportions of total score variance for each variance component. For the EPT writing, the proportions of total variance accounted for by persons, raters, and errors are 57.9%, 0% and 42.1% based on the rating method, and their counterparts are 55.9% 2.0% and 42.1% based on the rater method.

From a substantive perspective, the G-study results indicate that the largest proportion of total score variance is explained by true differences among examinee writing ability for the EPT writing. In addition, the estimated variance component for ratings/raters tends to be very small, suggesting that differences in rater severity/leniency are negligible. However, a sizeable error component is observed in the EPT writing (42.1%), suggesting the presence of person-by-rater interactions and/or other unidentified variability. Although the error component is nontrivial for the EPT writing, it does not considerably impact the precision of scores awarded to examinees when two ratings/raters or more are used, as will be discussed later in terms of SEMs.

Figure 2.2 presents score dependability and SEMs with respect to the number of ratings/raters from the writing component of EPT. Phi-coefficients indicate the extent to which awarded scores are reliable, while SEMs offer a different but relevant piece of information about the degree to which uncertainty exists in awarded scores. In Figure 2.2, the solid line represents the rating method, and the dash line refers to the rater method. Notably, phi-coefficients increase when more ratings/raters are used, whereas SEMs decrease as the number of ratings/raters increase. From a substantive perspective, Figure 2.2 (a) shows that for the EPT writing, the increase in phi-coefficients is larger when the number of ratings/raters changes from one to two, but the improvement is less dramatic when three ratings/raters or more are used. In addition, the results suggest that at least two raters/ratings are required to achieve a dependability of .70 or

higher for the EPT writing—this may be acceptable given the low-stakes of EPT as a university English placement test. Regarding the precision of awarded scores, when two ratings/raters are used, the SEM is expected to be .26 in Figure 2.2 (b). This converts to 1.04 points with 95% confidence interval (equivalent to four SEMs here) and ensures that uncertainty in awarded scores is not likely to be greater than 1.04 scale levels, which is acceptable for the EPT writing scale of 2-4. In sum, taking phi-coefficients and SEMs into consideration, two ratings/raters or more are needed for EPT operational use.

From a methodological perspective, results indicate that the rating and rater methods yield similar results in terms of reliability estimates and measurement errors in the EPT writing. With respect to the number of ratings/raters, phi-coefficients based on the rating method are consistently higher than those based on the rater method in Figure 2.2 (a), but the differences are negligible. SEMs, conditional on the number of ratings/raters, in Figure 2.2 (b) are almost identical with respect to the two methods.

2.6 Discussion

By means of simulation and empirical studies, the current paper reports on findings from comparing two ANOVA-based methods designed to handle sparse rated data under the G-theory framework. It demonstrates how methodological approaches to be applied to empirical research can be informed by simulation research. Depending on the particular compositions of score variability among raters, some estimated variance components may be different based on the rating and rater methods. Moreover, simulation results suggest that increasing the number of examinees improves the precision in estimating variance components, but the improvement is relatively small as the number of ratings/raters increases. These observations are congruent with

the general concept about G-study estimates that estimation precision is expected to improve by increasing the objects of measurement but not the elements within a facet (Brennan, 2001).

In this paper, different realistic rater scenarios were considered in the simulation study, and therefore recommendations for analytical approaches to handling sparse datasets in practice are discussed in light of these rater scenarios. First, when raters are assumed to be matched for their experience and/or training in a rated test, either the rating or the rater method yields good estimates of variance components. These results, together with the rating method requiring less computational complexity, suggest that the rating method is recommended in practice when raters can be expected to exhibit similar variability in their scoring, analogous to the assumption of homogeneous score variances across raters.

Second, when some raters are assumed to have more score variability than the rest, the estimated variance components for the person effect ($\hat{\sigma}_p^2$) and for the error component ($\hat{\sigma}_e^2$) are generally accurate based on both the rating and rater methods. Nonetheless, the estimated variance component $\hat{\sigma}_r^2$ based on the rating method tends to underestimate its true parameter σ_r^2, whereas the rater method seems to be more precise in this case. As a result, the rater method is recommended in operational use when a mixture of novice and seasoned raters participate together in scoring.

Empirical results from the writing component of EPT in this study suggest that two ratings/raters are sufficient to achieve acceptable score dependability and to control for measurement errors in operational use. In addition, the empirical study shows that the rating and rater methods are comparable in terms of estimating variance components, score dependability, and SEMs. This could be a result of the EPT raters being well trained as they have studied the rating rubrics intensively and have practiced scoring sample essays before scoring the

operational tests. These empirical results resonate with the simulation results that both methods produce good estimates when raters are matched for their training. Hence, given the comparability of estimation results between the two methods in the current empirical study, the rating method—a computationally less complex method—is sufficient for operational use, provided that raters are well trained and calibrated. However, it should be noted that when raters are well-trained in a rated test, the true parameter σ_r^2 can be very small and perhaps close to zero. In such cases, even a slight underestimation by the rating method can result in a negative $\hat{\sigma}_r^2$, which is at odds with the notion of variance component. This could be an explanation for the observed negative estimated variance component of ratings (-.0022) from the EPT data based on the rating method.

Finally, it is important to point out that the impact of underestimating σ_r^2 by the rating method is not large in terms of computing score reliability and SEMs in this paper because the true parameter σ_r^2 is small, compared to the other two true parameters involved in the computations—see Equations (2.2) and (2.3) for phi-coefficient and SEM. The true variance component σ_r^2 in this paper is empirically derived from writing studies with well-trained raters (Elorbany & Huang, 2012; Gebril, 2009; Lee & Kantor, 2005, 2007), and thus its value tends to be small. However, this may not always be the case in operational settings. If the relative magnitude of σ_r^2 is large compared to other variance components, the impact of its underestimation can be substantial. Huang and Foote (2010) investigated reliability of writing scores given by professors to nonnative speakers of English in a university setting. These professors were from various disciplines and did not have much experience and training in grading essays written by nonnative speakers of English. The authors reported that $\hat{\sigma}_p^2$, $\hat{\sigma}_r^2$ and $\hat{\sigma}_e^2$ were .056 (4% of total variance), .728 (52% of total variance) and .615 (44% of total

variance), respectively. In this particular case, the relative proportion of $\hat{\sigma}_r^2$ is high, and if these empirical values are to be taken as the true parameters, the underestimation of σ_r^2 by the rating method can unduly inflate score reliability, leading to possible false claims about score reliability.

2.7 Final Remarks

For the purpose of estimating variance components from sparse rated data, the results reported in this paper suggest that the rater method is preferred over the rating method when raters are expected to vary in their score variability due to potential differences in their rating experience and training. On the other hand, the rating method is recommended for the sake of simplicity when raters are matched for these background variables. A follow-up question for future research is whether statistical tests can be used or developed to identify heterogeneous rater variability and thus complement the recommendations presented here. For example, the Levene's test for equal variances from different raters (Levene, 1960) or the Mauchly's test for equal variances of differences between all possible pairs of raters (Mauchly, 1940) may be good candidates.

The current paper is limited to comparisons between two ANOVA-based methods of handling sparse rated data under the G-theory framework. Future research can compare these methods with other methods in dealing with sparse rated data (see Schoonen, 2005, for an application of structural equation modeling). Additionally, this paper is based on a one-facet random effect model in which raters were treated as a random facet. The current study can be further expanded to include writing tasks as a random facet since the most common facets involved in performance-based language assessments are those associated with tasks and raters (Lee, 2006). Another extension of the current study can be in the context of speaking assessments. The true parameters of variance components in this paper were derived from a

number of empirical writing studies; however, the relative proportions of variance components in writing assessments may be different from those in speaking assessments, and therefore the degree of underestimating variance components observed in this paper needs to be further investigated in the speaking-assessment context.

CHAPTER 3

ENGLISH LANGUAGE PROFICIENCY (ELP) STANDARDS-TO-STANDARDS CORRESPONDENCE RESEARCH: REVIEWER CONSISTENCY

Assessment-to-standards alignment in a standard-based assessment system typically can be evaluated by conducting item reviews, during which a panel of content experts looks for matches between test items and target content standards (Sireci, 1998). In the context of assessing the content knowledge of English language learners (ELLs) under the No Child Left Behind Act (2002), the notion of content alignment has been broadened to include correspondence between English language proficiency (ELP) standards and academic content standards. Clearly, proficiency in disciplinary language, defined here as aspects of language use closely tied to academic disciplines, is relevant to ELLs learning content knowledge such as in mathematics.

Several alignment methodologies exist (see Martone & Sireci, 2009 for a comprehensive overview). However, as pointed out by Bailey, Butler, and Sato (2007), correspondence methodologies are relatively new and less widespread. In addressing correspondence, the World-class Instructional Design and Assessment (WIDA) Consortium has been conducting ELP standards-to-standards correspondence studies, which examine language demands underlying ELP standards and academic content standards adopted by member states. Association between the two sets of standards can provide information about the extent to which ELLs are expected to encounter disciplinary language use that facilitates their content learning. Standards-to-standards correspondence, together with assessment-to-standards alignment, is therefore an integral part of

valid assessment-based inferences about ELL achievement levels in subject areas. In light of correspondence, the goals of the current study are to:

- describe one research methodology (see Bailey et al., 2007 for another approach) for studying the association between ELP standards and academic content standards; and
- more importantly, offer recommendations for future planning of ELP standards-to-standards correspondence studies based on empirical quantitative and qualitative results.

3.1 Context

The expanded view of content alignment for ELLs encompasses correspondence, which is partly motivated by the theoretical perspective that academic language can be conceptualized as a part of general language proficiency specifically associated with language use in subject areas (Adams, 2003; Schleppegrell, 2007) and by research recommendations that users of high-stakes academic achievement tests should consider the effect of English language proficiency on inferences made from test scores of ELLs (Butler, Orr, Gutiérrez, & Hakuta, 2000; see Solórzano, 2008 for a research synthesis). This expanded view is also reflected in the federal non-regulatory guidance in relation to ELP standards for Title III requirements:

> English language proficiency standards must, at a minimum, be linked to the State academic content and achievement standards. States are encouraged, but not required, to align English language proficiency standards with academic content and achievement standards. (U.S. Department of Education, Office of English Language Acquisition, 2003, p.9)

Thus, correspondence reflects both theoretical and policy concerns for valid score interpretations for the ELL population in a standard-based assessment system.

In addressing correspondence, Cook (2006, 2007) adapted Webb's (1997, 2007) alignment methodology to examine the association between ELP standards and state or school-district content standards. Strong association between the two sets of standards can ensure that ELLs are sufficiently exposed to academic language that complements their content learning. For example, for ELLs to understand and convey measurement concepts in comparing object lengths and sizes, their linguistic fluency in comparative forms (e.g., shorter than; larger than) must be developed.

3.2 The Current Study

In the current study, the WIDA English Language Proficiency Standards (World-class Instructional Design and Assessment, 2007), henceforth as the WIDA ELP Standards, were evaluated against the Common Core State Standards for Mathematics and English Language Arts (Common Core State Standards Initiative, 2010a, 2010b), henceforth as the CCSS. The former are performance standards which describe the degree to which ELLs can perform content-based linguistic tasks according to a language development continuum, whereas the latter are content standards which outline disciplinary knowledge to be learned.

The WIDA ELP Standards are organized by five grade-level clusters: preK-K, 1-2, 3-5, 6-8, and 9-12. Within each grade cluster, there are five subareas: Social and Instructional Language, Language of Language Arts, Language of Mathematics, Language of Science, and Language of Social Studies. Each subarea spans five language proficiency levels and covers four language domains: listening, speaking, reading, and writing. At the most fine-grained level, the model performance indicators (MPIs) serve as functional samples of language use in relation to ELLs reaching academic content expectations. For instance, the following MPI is taken from the Language of Mathematics in 6-8: *discuss how to solve problems using different types of line*

segments or angles from diagrams. It has three components: language function (i.e., *discuss*), content stem (i.e., *different types of line segments or angles*), and type of support (i.e., *diagrams*).

The CCSS for Mathematics are organized by grades, content domains, and standards. Content domains are larger categories that group related standards together, whereas standards delineate what students are expected to understand and to be able to do at the end of each grade. The CCSS for English Language Arts are organized by grades, strands, and anchor standards. Strands include reading, writing, listening and speaking, and language. Within each strand, anchor standards describe grade-appropriate content expectations.

3.3 Method

For each grade cluster examined in the current study, MPIs germane to the Language of Mathematics were evaluated against mathematics standards from the CCSS for Mathematics, covering the corresponding grade levels in that grade cluster. For example, the MPIs from the grade cluster 3-5 in the Language of Mathematics were paired with the Common Core mathematics standards in grades 3, 4, and 5. Likewise, MPIs in the language domain of reading were evaluated against the reading anchor standards from the CCSS for English Language Arts; writing MPIs were paired with the Common Core writing standards; speaking and listening MPIs combined were evaluated against the Common Core speaking and listening standards.

3.3.1 Data collection and participants. Data were collected from the study of correspondence between the WIDA ELP Standards and the CCSS for Mathematics and English Language Arts conducted in November, 2010. Forty-seven reviewers (N = 47) participated in the study, consisting of content teachers and ESL teachers from 18 WIDA member states. The reviewers were recruited and grouped by the five grade clusters. All of them had at least one year of experience working with the MPIs. Grouping reviewers by grade cluster followed the way

MPIs were organized; moreover, this would ensure that the reviewers were working with grade levels in which they specialized.

3.3.2 Research procedures. Prior to the study, the reviewers attended a workshop and received training on the overall objectives and procedures of the correspondence study. The author of this dissertation participated in planning the correspondence study, training the reviewers, and collecting the data. Regarding the research procedures, the WIDA ELP standards-to-standards correspondence study involves two parts (Cook & Wilmes, 2007). First, the reviewers were asked to individually determine the cognitive complexity of each content standard (i.e., the CCSS) using the four Depth of Knowledge (DOK) levels defined by Webb (see Webb, 2002 for DOK descriptors). DOK level 1 is the lowest level, representing low cognitive-demand processing such as simple recall of facts or formulaic language use. At DOK level 4, the learners are expected to extend their thinking such as synthesizing information into a new concept. Following that, the reviewers participated in a consensus discussion to agree on a final DOK level for each content standard. The consensus process can be viewed as reviewer calibration, during which the reviewers had the opportunity to build a common understanding of DOK levels.

Second, the reviewers were asked to rate the cognitive demand of each MPI by using the same DOK scale, and then to identify content standard(s), if any, that they believed to be most closely associated with the MPI. This part of the study was done individually by the reviewers and no collaboration was involved. This research design allows investigators to study three aspects of correspondence between standards: link, depth, and breadth. Link refers to the presence of (or lack thereof) connections between the content standards and language performance indicators. Depth involves the comparison of cognitive complexity between the

content expectations and associated language expectations. Breadth concerns the proportion of content standards adequately supported by language performance indicators. Each aspect has its associated statistic (Cook, 2007).

Individual reviewer DOK ratings of the content standards from CCSS and of the MPIs from WIDA ELP Standards were used in the present study to investigate reviewer reliability. Other measures, i.e., link and breadth, collected during the correspondence study were not discussed here because they were not of primary interest in this study.

3.3.3 Data analysis. Intraclass correlation coefficients (ICCs) (Shrout & Fleiss, 1979) were used in the current study to indicate the degree of consistency among panels of reviewers in their DOK ratings. The unit of analysis is grade levels: K, 1, 2, 3, 4, 5, 6, 7, 8, 9-10, and 11-12. Specifically, the measurement models are the random-effect model ICC (2, k) and the mixed-effect model ICC (3, k), where k refers to the number of reviewers. The difference between the two models is that the former treats the reviewers as random and therefore the generalizability is to a population of similarly qualified and trained reviewers, whereas the latter treats the reviewers as fixed and hence the derived inferences are limited to the reviewers recruited in the current study. Note that ICCs based on the mixed- and random-effect models are equivalent to reliability and dependability indices in the generalizability-theory framework, when no distinction is made between generalizability studies and decision studies (Kane & Brennan, 1977).

At the end of the correspondence study, the reviewers responded to a questionnaire, asking them for their qualitative feedback on the two sets of target standards and the relationship between them. A list of open-ended questions was presented in the questionnaire. Only responses

to the question regarding the level of specificity in the target standards were analyzed in the current study. Responses to the other questions are to be analyzed in future studies.

3.4 Results

3.4.1 Reviewer reliability. Table 3.1 shows the ICCs regarding reviewer consistency in rating the DOK levels of CCSS. ICCs(3,k) are presented in parenthesis. Regarding the content standards from the CCSS for Mathematics, reviewer consistency was generally medium and high except in Grades 5, 6, *Number and Quantity*, and *Functions*. As for the reading content standards from the CCSS for English Language Arts, reviewer consistency was mostly medium and high except in Grades 6 and 9-10. The reviewers were consistent with their ratings of the writing content standards except in grades K, 1, and 6. Finally, reviewer consistency in judging the speaking and listening content standards was generally medium and high except in Grades 1 and 8.

Table 3.2 presents the ICCs of reviewer DOK ratings of WIDA MPIs. ICCs(3,k) are shown in parenthesis. For the MPIs related to the Language of Mathematics, the reviewers exhibited high consistency in their DOK ratings in all grades. Regarding the MPIs pertaining to the language domain of reading, reviewer consistency was generally medium and high expect in Grades K and 1. The reviewers were mostly consistent with their ratings of the writing MPIs. Finally, reviewer consistency in judging the speaking and listening MPIs was high in all grades.

3.4.2 Reviewer feedback. Three common themes emerged from reviewer responses to the question about the level of specificity in the target standards. Some proposed providing explicit examples to help interpret the standards:

> *Example 1.* I believe teachers will need to be encouraged to look at examples provided by the overview to help interpret the [content] standards. *Grade cluster K-2*.

Example 2. Providing examples within the MPIs is very helpful to interpret the [language] standard. *Grade cluster K-2*.

Others voiced their concerns about the level of specificity regarding the target standards:

Example 3. There is a lot of room for individual interpretation. More specificity to the wording and verbs used is needed. *Grade clusters 6-8 and 9-12*.

Example 4. The writing standards were much easier to align because the tasks were pretty specific. *Grade cluster 3-5*.

Still others thought about the effect of grain-size on instructions:

Example 5. Standards and assessments can be very helpful to drive instruction. These [content] standards seem too "fuzzy" to help drive instruction. *Grade cluster 3-5*.

3.5 Discussion

It is noteworthy that the reviewers were more consistent with their judgments about the MPIs than about the content standards. This could be attributed to reviewers having prior experience working with the MPIs; however, at the time of the correspondence study, the CCSS were relatively new to the reviewers, which could lead to inconsistency in their understanding of what the content standards entailed. Reviewer feedback Examples 3 and 5 echoed the low ICCs observed in mathematics in some grades. These observations suggest a familiarity effect of target standards on reviewer reliability in their DOK ratings, which highlights the importance of careful planning in an ELP standards-to-standards correspondence study. Before the introduction of CCSS, the familiarity effect was less obvious in WIDA ELP correspondence studies in that the studies then involved a state's content standards, with which the reviewers were relatively more familiar. Separate analyses were conducted to examine reviewer reliability in their DOK ratings from past WIDA ELP correspondence studies using state content standards. Results from these

analyses have shown that the ICCs were generally higher than those observed in this study with respect to reviewer DOK ratings of content standards in mathematics and reading.

Thus, prior to conducting an ELP correspondence study, it is crucial that there is ample time for the target standards and stake holders (i.e., students, teachers, parents, etc) to "simmer" together. Since ELP correspondence studies rely heavily on reviewer judgments, investigators should strive to identify reviewers who have sufficient training and experience working with the target standards so that reviewer reliability can be better assured.

Another point to be mindful of is the balance between content experts and ESL specialists recruited in an ELP correspondence study. The current study made an effort to evenly team up content teachers with ESL teachers. The decision was informed by research suggesting that the collaboration between ESL and content teachers in identifying common language tasks required in content courses contributed to both teacher and ELL awareness of academic language inherent in subject areas (Lee, LeRoy, Adamson, Maerten-Rivera, Thornton, & Lewis, 2008; Stoddart, Pinnal, Latzke, & Canaday, 2002).

Moreover, according to reviewer qualitative feedback, high level of specificity in the target standards helped boost their congruence in interpreting the standards (i.e., Examples 1, 2 and 3). This is not a trivial matter in that valid inferences about correspondence between standards rest on the extent to which the reviewers comprehend the standards in a similar fashion. If the gaps were wide, valid claims could not be made regarding correspondence. Hence, the recommendation is that given the target standards, standards-to-standards correspondence studies should be conducted at the highest possible level of specificity, which corresponds to the content standards and MPIs in the current study. Nonetheless, it is recognized that the more fine-grained the level is, the greater the workload will be for the reviewers. Practical and structural difficulties

may prohibit the highest level of specificity to be targeted; in such cases, providing explicit examples along with the less fine-grained descriptors is recommended to help bring reviewer interpretations of the target descriptors to a similar level.

The issue of variability in interpreting descriptors is not limited to the standards investigated in the current study. For the purpose of achieving a shared understanding of target standards among reviewers, research efforts that involve reviewer judgments about any performance and/or content descriptors could benefit from the implications presented here about aiming at the most fine-gained level of specificity in the target standards and providing explicit examples of what the standards are about. In addition, in many English as a foreign language (EFL) contexts, English language education is tied to English language learning guidelines at the national level. In such non-US contexts, intended stakeholders' familiarity with and understanding of the target guidelines are important matters to be considered if consistent implementation of the guidelines is sought.

Lastly, while the introduction of the CCSS allows most states now to have common academic content standards, similar efforts in streamlining ELP standards across states have been proposed at the framework level by Bailey and Wolf (2012). A common framework for developing ELP standards is promising in that it could bring potential benefits of having different sets of ELP standards targeting core language skills across academic disciplines. However, given the lack of agreement in the literature regarding the definition of academic language (Anstrom, DiCerbo, Butler, Katz, Millet, & Rivera, 2010), and more importantly, that the construct of language proficiency has been shown to be represented differently in currently available ELP standards (Wolf, Farnsworth, & Herman, 2008), it is possible that ELP standards derived from the same framework may take different theoretical perspectives in conceptualizing

the expected language loads in academic disciplines. Hence, while a common framework for ELP standards is a good attempt to level the field of ELP-standard development, the merit of conducting ELP standards-to-standards correspondence studies still holds with the advent of the common ELP framework.

3.6 Final Remarks

Reviewer reliability in this study was interpreted primarily on a descriptive basis. Future studies can take an inferential stance and investigate the number of reviewers needed to achieve desirable reliability in ELP standards-to-standards correspondence studies. More specifically, the generalizability theory (G theory) has been promising in assessing the reliability of assessment-to-standards results (e.g., Porter, Polikoff, Zeidner, & Smithson, 2008). By applying the G theory in the context of ELP correspondence studies, one can evaluate the reliability of reviewer judgments with respect to the number of reviewers, from which empirical suggestions about the number of reviewers to be recruited can be derived.

In addition, ELP correspondence was examined only at the standard level in the current study. Equally important in a standard-based assessment system are the assessments derived from the target standards. Thus, future research can also look into the correspondence between ELP assessment and academic content assessment to arrive at a more comprehensive view of the degree to which academic language demands for ELLs are adequately addressed in a standard-based assessment system.

CHAPTER 4

DEPENDABILITY OF REVIEWER JUDGMENTS ABOUT LANGUAGE PERFORMANCE INDICATORS: HOW MANY REVIEWERS?

In a standard-based assessment system, standards serve to promote cohesion such that achievement tests and instructional practices are aligned with expected learning outcomes (La Marca, Redfield,Winter, & Despriet, 2000; Porter, 2002; Webb, 1997). When the notion of standard-based content alignment is applied in the context of assessing content knowledge, such as mathematics and science, of English language learners (ELLs), one additional element must be addressed—the association between English language proficiency and academic content knowledge. Clearly, access to content knowledge requires some mastery of academic English language, which is defined as aspects of language that are closely associated with ELLs learning content knowledge (Anstrom, DiCerbo, Butler, Katz, Millet, & Rivera, 2010).

In addressing content alignment for ELLs, the World-class Instructional Design and Assessment (WIDA) Consortium (readers are directed to http://wida.us/ for more information about the WIDA Consortium) has been conducting English language proficiency (ELP) standards-to-standards correspondence studies, which examine language demands shared by ELP standards and academic content standards adopted by member states. Correspondence between the two sets of standards provides information about the extent to which ELLs are expected to encounter academic language use that facilitates their access to content knowledge. This piece of information, in conjunction with assessment-to-standards alignment, serves to contribute evidence to validity arguments in a standard-based assessment system, which is in accord with the modern paradigm of test validation (Kane, 2006). Thus, together with alignment,

correspondence is an integral part of valid assessment-based inferences about ELLs' achievement levels.

Given that reviewer judgments are most often used in ELP standards-to-standards correspondence studies, one area of significant importance is the reliability of their judgments about the language demands underlying ELP standards and academic content standards. As such, this paper demonstrates new applications of generalizability theory (Brennan, 2001; Cronbach, Gleser, Nanda, & Rajaratnam, 1972) in the context of examining correspondence between the two sets of standards based on reviewer judgments. Moreover, the current study follows up on research reported in Chapter 3 in offering practical recommendations for conducing correspondence studies. In particular, this study aims to investigate the number of well-trained reviewers needed to achieve an acceptable level of reliability and to control for a reasonable amount of measurement errors in their judgments. Note that the reviewers recruited in this study have had extensive experiences working with the target standards, and as a part of the study, they received rigorous training on identifying the association between language proficiency standards and content standards. Results can be used as references for planning future correspondence studies and can also be used to examine whether past studies have recruited sufficient number of reviewers.

4.1 English Language Proficiency (ELP) Standards-to-Standards Correspondence Studies

The notion of content alignment is not new in a standard-based assessment system. Typically, it can be done by conducting item reviews, during which a panel of content experts looks for matches between test items and target content standards (Sireci, 1998). A number of alignment methodologies exist (see CCSSO, 2002). Rothman, Slattery, Vranek, and Resnick, (2002) illustrated the Achieve assessment-to-standards alignment protocol developed by Achieve,

Inc. The Surveys of Enacted Curriculum (SEC) was originally developed to examine the alignment between content instructions and instructional materials (Porter, 2002), and later has been expanded to include alignment between standards, assessments, textbooks and classroom instructions (Porter, Smithson, Blank, & Zeidner, 2007). Webb's (1997, 2002, 2007) approach collects both qualitative and quantitative information about the alignment between standards and assessments. A comprehensive overview and comparison of the above three different alignment methodologies can be found in Bhola, Impar, and Buckendahl (2003), and Martone and Sireci (2009).

In the context of assessing ELLs' content knowledge under the No Child Left Behind Act (2002), the notion of content alignment has been expanded to include not only assessment-to-standards alignment, but also the correspondence between ELP standards and academic content standards. This is partly motivated by the current theoretical view that academic language can be conceptualized as a part of general language proficiency specifically tied to disciplinary language use (Bailey & Butler, 2004; Bailey & Huang, 2011; Bailey, Butler, & Sato, 2007; Cummins, 1980; Schleppegrell, 2004) and by research recommendations that users of high-stakes academic achievement tests should take into consideration the effect of English language proficiency on inferences made from test scores of ELLs (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 1999; Butler, Orr, Gutiérrez, & Hakuta, 2000; Kopriva, 2000; see also Solórzano, 2008 for a research synthesis). This expanded view of content alignment for ELLs is also reflected in the federal non-regulatory guidance related to ELP standards for Title III requirements (U.S. Department of Education, Office of English Language Acquisition, February 2003). Thus, together with alignment, correspondence reflects both research and policy concerns for valid score

interpretations of the ELL populations. In addressing ELP standards-to-standards correspondence, Cook (2006) adapted Webb's (1997) alignment methodology to examine the association between expected academic language loads for ELLs and academic content expectations (i.e., content standards). For example, for ELLs to understand and convey measurement concepts in comparing object lengths and sizes, their linguistic fluency in comparative forms (e.g., shorter than; larger than) must be developed.

4.2 The Current Study

In an ELP standards-to-standards correspondence study, a set of ELP standards for ELLs is paired with a set of academic content standards. With regards to the current study, the WIDA English Language Proficiency Standards (World-class Instructional Design and Assessment, 2007) are performance standards which describe the degree to which ELLs can perform content-based linguistic tasks according to a language development continuum, whereas a state's or nation's academic content standards (e.g., Common Core State Standards for Mathematics) outline knowledge to be learned in subject areas. Strong association between the two sets of standards suggests that ELLs are sufficiently exposed to disciplinary language use that complements their content learning. The degree of association is worth investigating in that English academic language is considered to be a significant factor in the academic success of ELLs (Francis, Rivera, Lesaux, Kieffer, & Rivera, 2006; Gottlieb, 2006).

The summative framework of the WIDA English Language Proficiency Standards is used in a WIDA ELP standards-to-standards correspondence study. It is organized by five grade-level clusters: preK-K, 1-2, 3-5, 6-8, and 9-12. Within each grade cluster, the Standards include five subareas: Social and Instructional Language, the Language of Language Arts, the Language of Mathematics, the Language of Science, and the Language of Social Studies. Each subarea spans

five language proficiency levels and covers four language domains: listening, speaking, reading, and writing. The five language proficiency levels are characterized by a developmental progression in the complexity of language use encountered and produced by the ELLs. For instance, the language proficiency at level 1 is limited to understanding pictorial or graphical representations of general language related to academic disciplines. ELLs at level 5 are expected to process and use specialized or technical disciplinary language.

At the most fine-grained level, the model performance indicators (MPIs) are defined at the combination of each language domain and proficiency level (see Appendix A for an example matrix of MPIs). MPIs serve as functional samples of language use in relation to ELLs reaching academic content expectations. For example, the following MPI is taken from the Language of Mathematics: *match vocabulary associated with perimeter or area with graphics, symbols or figures*. It has three components: language function (i.e., *match*), content stem (i.e., *perimeter or area*), and type of support (i.e., *graphics, symbols or figures*).

In a WIDA ELP standards-to-standards correspondence study, panels of reviewers were asked to rate the cognitive complexity of each MPI, and then to identify descriptor(s), if any, from the academic content standards that they believe to be most closely associated with the MPI. Researchers (Webb, Herman, & Webb, 2007; Porter, Polikoff, Zeidner, & Smithson, 2008) acknowledged the importance of investigating rater reliability in assessment-to-standards alignment studies, and given that the procedures of WIDA ELP correspondence studies are directly derived from Webb's (1997) alignment procedures, reviewer reliability is also important in correspondence research and needs to be examined so that inferences about the relationship between the two sets of standards can be more confidently drawn.

Although a few studies have examined the effect of the number of reviewers on their reliability in judging alignment between assessments and standards (e.g., Herman, Webb & Zuniga, 2007; Porter et al., 2008), no study has looked into this topic in ELP standards-to-standards correspondence studies. Furthermore, while the use of generalizability theory in investigating reliability has been successful in studies involving human raters, such as performance-based assessments, (e.g., Gao, Shavelson, & Baxter, 1994; Gebril, 2009; Lee & Kantor, 2007; Shavelson, Baxter, & Gao, 1993; Xi, 2007), few applications of generalizability theory can be found in correspondence research, as pointed out by Martone and Sireci (2009). The current study seeks to advance applications of generalizability theory in a new research venue for ELP standards-to-standards correspondence studies and to fill the gap by evaluating the reliability and measurement errors of reviewer judgments with respect to the number of reviewers. The following two research questions are addressed:

1. What proportions of total rating variance can be attributed to variability in cognitive demands of MPIs and to variability among reviewers?
2. How many reviewers are required to ensure satisfactory reliability and to control for reasonable measurement errors in their judgments?

4.3 Method

4.3.1 Instruments and study participants. Data were collected from WIDA ELP standards-to-standards correspondence studies conducted from Fall 2007 to Fall 2011. For each correspondence study, the WIDA English Language Proficiency Standards (2007 edition) was paired with either a member state's or nation's academic content standards in subject areas of mathematics, language arts, or science. Table 4.1 provides a summary of the correspondence studies included in the current study.

As a part of the WIDA ELP standards-to-standards correspondence study, reviewers are asked to rate the cognitive complexity of MPIs using the four Depth of Knowledge (DOK) levels defined by Webb (see Webb, 2002 for DOK descriptors). Level 1 (Recall and Reproduction) is the lowest level, representing low cognitive-demand processing, such as simple recall of facts or formulaic language use. At DOK level 2 (Skills and Concepts), learners engage in two-step mental processing beyond a rote response. DOK level 3 (Strategic Thinking) requires abstract and high-level thinking. At DOK level 4 (Extended Thinking), the learners are expected to extend their thinking, such as synthesizing information into a new concept. In the current study, examination of reviewer reliability is based on their judgments about DOK levels of MPIs.

Generally, reviewers recruited in the correspondence studies consisted of ESL specialists and content teachers from the member states, and they were grouped by the five grade clusters. Recruiting reviewers from the member states ensures that they are familiar with both sets of standards examined in the correspondence study. Grouping reviewers by grade clusters follows the way MPIs are organized and ensures that the reviewers are working with grade levels with which they are familiar. In total, 327 reviewers participated in the 11 mathematics correspondence studies, 216 reviewers in the 8 language arts studies, and 235 reviewers in the 9 science studies. Note that for some correspondence studies in science, not all grade clusters were studied.

4.3.2 Data collection procedures. Data collection procedures in a WIDA ELP standards-to-standards correspondence study were adapted from Webb (1997). All data were collected via the Web Alignment Tool (Webb, Alt, Ely, Cormier, & Vesperman, 2005), an online tool originally designed to collect and analyze assessment-to-standards alignment data. However, for the purpose of WIDA ELP correspondence studies, the online tool was used for collecting data,

not for analyzing. A brief discussion of the WIDA ELP correspondence study procedures is provided next; however, full details are not discussed here as the scope of this study focuses on the reliability of reviewer judgments in WIDA ELP correspondence studies. Readers interested in full procedures of WIDA ELP correspondence studies are directed to a report prepared by Cook and Wilmes (2007).

A correspondence study involves two parts. First, reviewers attended a workshop and received intensive training on the overall objectives and procedures of WIDA ELP correspondence study. The reviewers then participated in a consensus discussion to determine the DOK levels of academic content standards. This part of the study can be viewed as reviewer calibration, during which the reviewers had the opportunity to become familiar with the online-tool functionalities and to develop a common understanding of DOK levels. Second, all reviewers were asked to associate each WIDA MPI with content descriptor(s), if any, from the academic content standards, and to determine the DOK level of each MPI. This part of the study was done independently by each reviewer and therefore no collaboration was involved. The author of this dissertation was responsible for setting up the correspondence studies, training the reviewers, and collecting the data using the online tool (i.e., the Web Alignment Tool). Individual reviewer DOK ratings of MPIs were used in the analyses to investigate the reliability of their judgments in WIDA ELP correspondence studies. Other measures collected in the correspondence study were not of primary interest and therefore not included here.

4.3.3 Data analysis. Generalizability theory (G theory) is a random facet measurement model which conceptualizes observed scores as a composite of various sources, or *facets*, in addition to the objects of measurement (Brennan, 2001; Cronbach, Gleser, Nanda, & Rajaratnam, 1972). In classical test theory, an observed score can be broken down into a true score and a

single error term. G theory liberates classical test theory via certain analysis of variance (ANOVA) procedures in the sense that the single error term can be further decomposed into variance components corresponding to the various facets in measurement.

Conceptually, G theory distinguishes between generalizability studies (G studies) and decision studies (D studies). G studies estimate variance components of the main and interaction effects relevant to the objects of measurement and facet(s), whereas D studies use these estimates for decision making in terms of effective measurement designs (Brennan, 1992). Take the current study for example. The objects of measurement are MPIs (measured on the DOK scale) and the random facet is reviewers; thus, effective measurement procedures are examined by varying the number of reviewers in D studies.

Since groups of reviewers assigned DOK levels to MPIs organized by the five grade clusters, the unit of analysis is grade cluster in the current study. Individual G studies were first performed for each grade-cluster group and these G-study results were then averaged across correspondence studies to obtain averages of estimated variance components. The decision to analyze and aggregate individual G-study results by grade clusters also follows how the reviewers were recruited in operational studies.

Notations and equations presented here follows those from Shavelson and Webb (1991). In general, the G studies performed in this study employ a two-facet $m \times r \times g$ mixed effect model, where m is the objects of measurement (i.e., MPIs), r refers to the random reviewer facet and g is the fixed grade-level facet for each grade cluster. Depending on the number of grades examined in each grade cluster, the G study follows a one-facet fully crossed $m \times r$ random effect model when the number of grade is one; otherwise, it becomes a two-facet fully crossed $m \times r \times g$ mixed effect model when the number of grades is at least two. A total of 136

individual grade-cluster G studies were conducted in this study, including 55 in mathematics, 40 in language arts, and 41 in science. Table 4.2 provides the number of random- and mixed-model G studies. In either case, the random portion of the study design was MPIs (*m*) fully crossed with reviewers (*r*), and therefore the variance components to be estimated are MPIs, reviewers, and the interaction between the two plus other unidentified variability and random errors.

Computations of the estimated variance components are discussed next. The total rating variance of MPIs in the $m \times r$ random effect model can be expressed by $\sigma^2(X_{mr}) = \sigma_m^2 + \sigma_r^2 + \sigma_{mr,e}^2$, whereas the total rating variance in the $m \times r \times g$ mixed effect model can be decomposed into the following seven variance components: $\sigma^2(X_{mrg}) = \sigma_m^2 + \sigma_r^2 + \sigma_g^2 + \sigma_{mr}^2 + \sigma_{rg}^2 + \sigma_{mg}^2 + \sigma_{mrg,e}^2$. Estimated variance components of the $m \times r$ random effect model can be directly derived from sample mean squares (see Shavelson & Webb, 1991, p. 28 for details). Estimated variance components of the $m \times r \times g$ mixed effect model are also derived from sample mean squares but are averaged over the fixed grades in each grade cluster as follows:

$$\begin{aligned} \hat{\sigma}_{m*}^2 &= \hat{\sigma}_m^2 + \hat{\sigma}_{mg}^2/n_g, \\ \hat{\sigma}_{r*}^2 &= \hat{\sigma}_r^2 + \hat{\sigma}_{rg}^2/n_g, \text{ and} \\ \hat{\sigma}_{mr,e*}^2 &= \hat{\sigma}_{mr}^2 + \hat{\sigma}_{mrg,e}^2/n_g, \end{aligned} \tag{4.1}$$

where n_g is the number of fixed grades in each grade cluster. The computer program GENOVA (Crick & Brennan, 1982) was used in estimating variance components of the main and interaction effects relevant to MPIs, reviewers and grades (if applicable). Since the program does not handle fixed effects, variance components of the mixed-effect models were estimated as if the models were random and then these estimates were entered into Equation 4.1 to obtain the mixed-effect estimates.

Next, estimated variance components for each grade cluster were averaged across correspondence studies in each content area given that average estimates are more stable than estimates from individual correspondence studies and that the derived inferences are for a general correspondence study rather than a specific one. Consequently, D studies were performed based on these average estimates. The purpose of D study is to optimize measurement procedures so that a desirable reliability-like coefficient can be achieved and that measurement errors can be controlled (Brennan, 2000). A total of 80 D studies (8 individual D studies for each of the five grade clusters per content area) were conducted in the current study by increasing the number of reviewers from one to eight. In particular, the absolute interpretations were adopted because the DOK scale on which the MPIs were measured was criterion-based, describing the levels of cognitive demand associated with the performance indicators. Table 4.3 presents the formulas for calculating absolute error variance, standard error of measurement and phi-coefficient for D-study $m \times R$ random effect design and $m \times R \times G$ mixed effect design, where $n_r^{'}$ refers to the number of reviewers in D studies. By varying the number of reviewers, the effect of number of reviewers on phi-coefficients and standard errors of measurement in reviewer DOK ratings of MPIs can be studied.

4.4 Results

4.4.1 Proportions of total variance attributed to MPIs and reviewers. For mathematics, 11 individual G studies were conducted per grade cluster. Table 4.4 shows the average estimates of variance components, their relative proportions, and their standard errors in the mathematics studies. N refers to the number of correspondence studies conducted in each grade cluster; thus, the standard error of an average estimated variance component is the standard deviation of the estimate divided by $\sqrt{\mathrm{N}}$.

Regarding the percentage of variance components for each grade cluster in mathematics, the variance component of MPIs has the lion share except for preK-K. The proportion of total rating variance attributed to reviewers is relatively small. The variance component of error term (*mr,e*) constitutes a sizeable portion of total variance. As for the variance components across grade clusters, the percentage of MPI variance component is the highest in 3-5 (63%) and lowest in preK-K (45%). Variability among reviewers is the highest in 1-2 (12%) and lowest in 9-12 (6%). PreK-K has the highest proportion of error term variance (47%) while 3-5 has the lowest (29%).

For language arts, 8 individual G studies were performed per grade cluster. The average estimates of variance components, their relative proportions and standard errors are presented in Table 4.5 by grade clusters. Within each grade cluster, the percentage of total rating variance due to MPIs is the largest except for preK-K. Reviewer variability constitutes the smallest proportion. The variance component of the error term accounts for the largest portion in preK-K and also exhibits considerable contributions in the other four grade clusters. Across the five grade clusters, variability among MPIs is the lowest in preK-K (22%) and the highest in 3-5 (78%). The proportions of reviewer and error term variance are the highest in preK-K and lowest in 3-5.

For science, 7 individual G studies each were conducted in grade clusters preK-K and 1-2, whereas 9 G studies each were performed in 3-5, 6-8, and 9-12. Table 4.6 presents the average estimates of variance components, their relative proportions, and their standard errors. For each grade cluster, it appears that the percentage of total variance accounted for by MPIs is the largest except for preK-K. The proportion of reviewer variance component is the smallest, contributing to less than 10% of the total variance. The percentage of error term variance is the largest in preK-K and is also substantial in the other grade clusters. Across the grade clusters, the

proportion of MPI variance component is the highest in 6-8 (71%) and lowest in preK-K (37%). Variability among reviewers is higher in low grade clusters than in high grade clusters. The percentage of error term variance is the highest in preK-K (54%), and exhibits similar contributions to the total variance in the other grade clusters, ranging from 26% to 33%.

4.4.2 Increasing the number of reviewers from one to eight. For mathematics, a total of 8 D studies were conducted for each grade cluster by increasing the number of reviewers from one to eight, amounting to 40 D studies. Figure 4.1 displays the phi-coefficients and standard errors of measurement (SEMs) in relation to the number of reviewers. Overall, the phi-coefficient of reviewer judgment is the highest in 3-5, followed by 6-8, 9-12, 1-2, and preK-K, but the differences between grade clusters 3-5, 6-8, and 9-12 are minimal. PreK-K has the largest measurement errors, followed by 9-12, 6-8, 1-2, and 3-5. The difference in SEM between 6-8 and 1-2 is negligible.

Figure 4.1 (a) shows a horizontal line representing the phi-coefficient at 0.6. Cook and Wilmes (2007) proposed that an intraclass correlation coefficient (ICC) of 0.7 or larger indicates good reliability among reviewers in standards-to-standards correspondence studies. Inherent in ICC are relative interpretations about reviewer judgments. Given that the current study focuses on absolute interpretations, the reliability-like coefficient based on absolute criterion can be lower than that based on relative criterion, and thus the minimal phi-coefficient is set at 0.6. It appears that a minimum of two reviewers is required to achieve a phi-coefficient of at least 0.6 in grade clusters preK-K, 1-2, and 9-12 while one reviewer is sufficient in 3-5 and 6-8. Figure 4.1 (b) has a horizontal line representing the SEM at 0.25. If an investigator wishes to control measurement errors in reviewer judgment so that the difference between the upper and lower 95% confidence limits is not greater than one (equivalent to four SEMs in this case) on a DOK

scale of 1-4, at least three reviewers are needed in 3-5, four reviewers in 1-2, 6-8 and 9-12, and five reviewers in preK-K.

For Language Arts, eight D studies were conducted in each grade cluster by increasing the number of reviewers from one to eight. Phi-coefficients and SEMs in relation to the number of reviewers are presented in Figure 4.2. The grade cluster 3-5 has the highest dependability index, followed by 6-8, 9-12, 1-2 and preK-K. On the other hand, the SEM is the largest in preK-K, followed by 9-12, 1-2, 6-8 and 3-5.

Figure 4.2 (a) suggests that one reviewer can achieve an acceptable level of dependability in 3-5 and 6-8 while two reviewers are sufficient in 1-2 and 9-12; nonetheless, at least six reviewers are required to reach a minimal phi-coefficient of 0.6 in preK-K. With respect to measurement errors, Figure 4.2 (b) shows that at least three reviewers are needed in 3-5, four reviewers in 1-2 and 6-8, and five reviewers in preK-K and 9-12 so that the 95% confidence limit of DOK rating is within one on the DOK scale of 1-4.

For science, eight D studies were conducted for each grade cluster by increasing the number of reviewers from one to eight, amounting to 40 D studies. Figure 4.3 shows the phi-coefficients and SEMs with respect to the number of reviewers. It can be observed that the grade cluster 6-8 has the highest phi-coefficient followed by 9-12, 3-5, 1-2, and preK-K, but the difference between 3-5 and 9-12 is minimal. As for measurement errors, preK-K has the largest SEM, followed by the other four grade clusters whose differences in SEMs are not greater than 0.04.

Figure 4.3 (a) shows that only one reviewer is sufficient to reach satisfactory phi-coefficient in 3-5, 6-8, and 9-12, whereas two reviewers are needed in 1-2 and three reviewers in preK-K. Figure 4.3 (b) suggests that at least three reviewers are needed in 1-2 to control for a

reasonable amount of measurement errors. A minimum of four reviewers is required in 3-5, 6-8 and 9-12, and six reviewers in preK-K.

4.5 Discussion

The current study demonstrated a systematic way of investigating reliability-like coefficients and measurement errors among a panel of reviewers when the objects of measurement were their judgments about the cognitive demands of ELL performance descriptors associated with disciplinary language use. More specifically, the current study sought to determine the number of reviewers needed for optimal measurement procedures.

4.5.1 Findings from G studies. A discussion of the proportion of total rating variance attributable to variability in the DOK levels of MPIs and to variability among reviewer judgments can provide insights into the relative cognitive complexity of MPIs and the degree to which the reviewers varied in their judgments. It is noteworthy that within each grade cluster, the percentage of total variance due to MPIs is the largest except for preK-K in the three content areas investigated in this study. One reason could be that the WIDA MPIs in preK-K were designed to target low cognitive processing, leading to less variability in their DOK levels. Alternatively, it might be the case that the DOK scale adopted in the current study was not refined enough to capture subtle differences in the cognitive demands of performance indicators for kindergarten ELLs. Nonetheless, the DOK scale adequately differentiated the cognitive complexity of MPIs in the other four grade clusters.

It is also noteworthy that the percentage of total variance attributed to the variability among reviewers is generally small. This could be explained by the fact that the reviewers have had experiences working with the WIDA English Language Proficiency Standards, and hence they shared a common understanding of what the MPIs entailed. In addition, the trivial

variability among reviewers in most grade clusters could be attributed to the effective workshop training and consensus discussion of DOK levels in the first part of correspondence studies.

Although the error-term variance component contributes to a sizeable proportion of total variance in each grade cluster, the interpretation of this variability is not conclusive in that the interaction between MPI and reviewer is confounded with unmeasured systematic variability and random errors. Substantial error-term variance might suggest that the rank ordering of MPI cognitive demands differs by the reviewers to a considerable degree. However, it could also be the case that the rank ordering varies minimally for different reviewers but unmeasured sources of variability are large.

4.5.2 Findings from D studies. Some considerations in interpreting D-study results are warranted before discussing practical implications regarding effective measurement procedures. Brennan, Gao, and Colton (1995) strongly argued that "standard errors of measurement are almost always more informative for decision making than are generalizability coefficients" (p.168). In some circumstances, an investigator may be more concerned with the precision of measurement and is willing to live with a somewhat lower reliability or vice versa. In other circumstances, the precision and reliability are equally crucial. Given that the current study sought to determine the minimal number of reviewers required to reach acceptable reliability and to control for reasonable measurement errors, the criteria adopted here to determine the number of reviewers were based on phi-coefficients and SEMs simultaneously.

For mathematics, the phi-coefficient suggests that two reviewers can achieve a satisfactory level of reliability in preK-K, while the SEM suggests that five reviewers are required to reduce measurement errors to a modest amount. As a result, a minimum of five reviewers are recommended in preK-K. In 1-2 and 9-12, two reviewers are suggested on

reliability grounds, while four are suggested on precision grounds—minimally four reviewers are recommended in 1-2 and 9-12. Only one reviewer is needed in 3-5 to reach a phi-coefficient of 0.6; however, three reviewers are required so that the range of measurement errors is not greater than one DOK level. Consequently, three reviewers are recommended in 3-5. From a reliability perspective, only one reviewer is needed in 6-8; nevertheless, four reviewers are required from a precision perspective, and hence four reviewers are recommended in 6-8.

For language arts, six reviewers are needed in preK-K to achieve a phi-coefficient of 0.6 while five reviewers are required so that the measurement errors are reasonable. Consequently, six reviewers are commended for preK-K. From a dependability perspective, two reviewers are needed in 1-2; nonetheless, four reviewers are required from a precision perspective, and therefore four reviewers are recommended in 1-2. The phi-coefficient suggests that only one reviewer is required in 3-5, while the SEM points to three reviewers and hence three reviewers are commended in 3-5. In 6-8, only one reviewer is suggested on dependability grounds, while four are suggested on precision grounds—minimally four reviewers are recommended in 6-8. The phi-coefficient suggests that two reviewers can achieve a satisfactory level of dependability in 9-12, while the SEM suggests that five reviewers are required to reduce the measurement errors to a modest degree. As a result, a minimum of five reviewers are recommended in 9-12.

For science, three reviewers are needed in preK-K from a reliability perspective; nonetheless, six reviewers are required from a precision perspective, and therefore six reviewers are recommended in preK-K. The phi-coefficient suggests that two reviewers are required in 1-2, while the SEM points to three reviewers; hence three reviewers are recommended in 1-2. In grade clusters 3-5, 6-8, and 9-12, only one reviewer is suggested on reliability grounds, while

four are suggested on precision grounds. As a result, four reviewers are recommended in 3-5, 6-8, and 9-12.

4.6 Final Remarks

4.6.1 Implications. Practical and structural difficulties may arise in operational studies. Thus, if resources permit, it would be ideal to recruit at least one more reviewer beyond the minimal recommendations here to account for unexpected absence from the reviewers. If not, an investigator may want to be flexible with the recommendations. That is, instead of considering reliability and measurement errors simultaneously, the investigator may want to look at recommendations based on phi-coefficient only, provided that reviewers participate in a well-designed training workshop prior to the actual study so that some control of measurement errors can be expected. This is based on the assumption that reviewer training is likely to reduce variability among reviewer judgments, leading to smaller SEM if all else being equal, as can be observed from the SEM equations in Table 4.3. Alternatively, the investigator can focus on SEM only provided that, in addition to training, the language descriptors are expected to vary in their levels of cognitive demand to some extent. This is based on the assumption that greater variability in the objects of measurement (σ_m^2) may result in a larger phi-coefficient if all else being equal, as can be observed from the phi-coefficient equation in Table 4.3. Table 4.7 presents recommendations for the minimal number of reviewers needed within each grade cluster in the content areas of mathematics, language arts, and science based on the three approaches discussed here: reliability only, precision only, and both.

4.6.2 Numbers of recommended reviewers by grade levels. Alternatively, an investigator may be interested in a specific grade level only and wishes to operate at a more fine-grained level than the grade clusters investigated in the previous sections. Additional analyses

were conducted at each grade level. As such, the grade-level G studies for each subject area follow a one-facet fully crossed $m \times r$ random effect model, where the MPIs were fully crossed with the raters. For each grade level per subject area, variance components were first estimated and then averaged across the total number of correspondence studies included in this study. Take 1st grade mathematics for instance. Variance components in relation to the MPIs, raters, and errors were estimated separately for the 11 mathematics correspondence studies. Next, these estimates were averaged across the 11 correspondence studies to arrive at the final variance-component estimates for computing phi-coefficients and SEMs. Table 4.8 presents recommendations for the minimal number of reviewers needed for each grade level in mathematics, language arts, and science ELP standards-to-standards correspondence studies. Note that for the grade clusters preK-K and 9-12, grade-level data were not available and therefore the recommendations shown in Table 4.8 for these two grade clusters are duplicates of the previous grade-cluster results from Table 4.7.

As can be observed, the grade-level analysis yields recommendations different from those based on the grade-cluster analysis. Take mathematics for example. By comparing Table 4.7 with Table 4.8, one can see that the numbers of reviewers recommended in grade levels 1-8 are consistently larger than their counterpart grade-cluster recommendations based on the three approaches adopted in this study. The sharpest contrast is observed in 3rd grade mathematics when SEM is used as the sole criterion. The grade-cluster analysis suggests 3 reviewers, whereas the grade-level analysis points to 6 reviewers. For language arts and science, almost all grade levels have a larger number of recommended reviewers than that based on the grade-cluster analysis. The differences in the recommended numbers of reviewers are expected. Due to the fact that the grade-level data are aggregated in the grade-cluster analysis, discrepancies among

reviewer judgments are expected to be leveled out to some extent in the grade-cluster analysis, resulting in higher reviewer reliability and subsequently fewer numbers of recommended reviewers based on the grade-cluster analysis.

4.6.3 Limitations. From a conceptual stance, the current study did not fully conform to the G-theory framework in that G-theory analyses in this study were based on data collected from existing measurement procedures. In ideal G-theory applications, data are collected for the purposes of examining psychometric properties of an instrument in its developmental form and of designing optimal measurement procedures (e.g., Brennan, Gao, & Colton, 1995; Lee & Kantor, 2007). In the current study, the data were collected from the existing measurement procedures in operational correspondence studies to inform decisions regarding the number of reviewers required. Nevertheless, this conceptual deviation does not render the analyses and results presented here less meaningful, as the conceptual distinction between developmental and existing measurement procedures is not often made explicit in many published G-theory applications given that the purpose of these studies (and the current study) was to validate or refine the existing measurement procedures (e.g., Gebril, 2009; Sudweeks, Reeve, & Bradshaw, 2004; Xi, 2007).

The instruments used in this study were MPIs from the WIDA English Language Proficiency Standards (2007 edition) and Webb's DOK descriptors. It would be necessary to revisit the number of reviewers recommended here if other language performance indicators and cognitive complexity scales are used to investigate the academic language loads residing in content standards. In addition, caution is warranted when generalizing the results to content areas other than mathematics, language arts, and science in that different subject matters may differ in their linguistic demands for ELLs.

CHAPTER 5

EVALUATING TUKEY'S SINGLE-DEGREE-FREEDOM TEST FOR DETECTING NONADDITIVITY IN RATED MEASUREMENT

Generalizability theory or G theory (Brennan, 2001; Cronbach, Gleser, Nanda, & Rajaratnam, 1972; Shavelson & Webb, 1991) conceptualizes observed score variability as a linear combination of the true variation in the objects of measurement and other variations as a result of different measurement sources that are anticipated by or of interest to an investigator. For instance, in an essay exam on some scientific knowledge for a group of students, the object of measurement is a student's knowledge in science, and a potential source of measurement variation (i.e., *facet* in G-theory terminology) is score variability introduced by different raters scoring the essays. Ideally, one would like to see true differences in students' scientific knowledge reflect observed score variability as much as possible, not differences among rater severity/leniency.

In addition to gauging how much observed score variability is explicable by different measurement facets, G theory has been widely applied in the analysis of score reliability of rated measurement in large-scale assessments (e.g., Brennan, 2000; Brennan, Gao, & Colton, 1995; Gao, Shavelson, & Baxter, 1994; Lee & Kantor, 2007; Shavelson, Baxter, & Gao, 1993) and in classroom-assessment contexts (e.g., Gebril, 2009; Huang & Foote, 2010; Sudweeks, Reeve, & Bradshaw, 2004). In most G-theory applications in rater-mediated measurement, examinee responses were evaluated against some criterion-based standards or descriptors; hence, the absolute interpretation of reliability is adopted, which is defined as the phi-coefficient in G-theory framework.

In a G-theory design, an observed score is a linear function of the main and interaction effects of persons, facet(s), and errors. For example, a ($p \times r$) design takes the following linear model:

$$X_{pr} = \mu + \alpha_p + \beta_r + \varepsilon_{pr,e}, \tag{5.1}$$

where the score (X_{pr}) of person p given by rater r is the sum of an overall mean (μ) and the three random effects pertaining to persons, raters, and errors, where $\alpha_p \sim N(0, \sigma_p^2)$, $\beta_r \sim N(0, \sigma_r^2)$, and $\varepsilon_{pr,e} \sim N(0, \sigma_e^2)$, respectively. The overall mean can be considered as the performance of an average person on the measured construct. The person effect reflects the relative standing of a random person from the intended population compared to the average person, while the rater effect corresponds to the relative standing of a random rater from the intended population of raters compared to a rater with average severity/leniency.

In a typical rated assessment, each person is repeatedly rated across some or all raters and is rated once per rater; as a result, the person-by-rater interaction is confounded with the random errors. In other words, neither the interaction nor the random errors can be assessed independently, as has been pointed out by Cronbach et al. (1972). In addition, both the interaction and random-error terms are subsumed under the error component ($\varepsilon_{pr,e}$). It is data with this single-observation-per-cell layout that is of concern to the current study. Current G theory assumes additivity such that "all effects in the model are uncorrelated" (Brennan, 2001, p. 23). An example of the additive assumption is that the three random effects of persons, raters, and errors in Model (5.1) are not correlated with one another. As will be shown later, when the assumption of additivity is met such that all the effects are uncorrelated, the two confounding terms (i.e., interaction and random errors) within the error component does not need to be assessed separately in the estimation of variance components. However, when some or all of the

three random effects in Model (5.1) are correlated, the model becomes nonadditive, and the confounding nature of the error component will introduce additional complications in estimating variance components because the random-error term now needs to be estimated independently of the interaction term.

Statistically, the distinction between additive and nonadditive models is of great importance in that formulas for variance component estimates differ depending on the nature of the models. However, the difference between additivity and of nonadditivity is not adequately discussed in the G-theory literature, and hence the identification of potential nonadditivity in data is typically overlooked in G-theory applications. This chapter attempts to advance such a discussion and at the same time highlights the importance of detecting nonadditivity in G-theory applications.

5.1 Nonadditivity

The discussion of nonadditivity has been noted in the literature of Analysis of Variance or ANOVA regarding data with a single-observation-per-cell design. (Davis, 2002; Myers, 1979; Scheffe, 1999). Myers (1979) alluded to the fact that accurate estimation of variance components cannot be achieved with the presence of nonadditivity. Tukey (1949) developed statistical procedures to detect nonadditivity in data. In view of the advantages of working with additive data, Anscombe and Tukey (1963) proposed procedures that transform nonadditive datasets into additive ones. As a caveat of data scrutiny, Scheffe (1999) suggested that it would be helpful to examine the observed variance component for errors—a relatively large value may suggest nonadditivity and/or a violation of other ANOVA assumption(s). Zhang and Lin (2013) introduced an additivity index, measuring the degree to which nonadditivity exists in data. The smaller the index is, the larger the magnitude of nonadditivity would be. By incorporating this

index into the G-theory framework, the authors also developed nonadditive G theory for one-facet measurement designs.

Given that the use of G-theory leans heavily on ANOVA techniques in estimating variance components and that data associated with G-theory applications usually follow the single-observation-per-cell layout, issues with nonadditivity should deserve more attention from G-theory users. In relation to Model (5.1), Table 5.1 illustrates the difference between additivity and nonadditivity by presenting formulas for estimated variance components in a one-facet additive G-theory model (see Shavelson & Webb, 1991, p.28) and for those in a one-facet nonadditive G-theory model (Zhang & Lin, 2013).

The difference between additive and nonadditive assumptions has implications for estimating variance components for the person effect. As such, in the additive model, the numerator of the estimated person variance component in Table 5.1 is subtracted by the error variance component which encompasses the interaction ($\hat{\sigma}^2_{pr}$) and random errors ($\hat{\sigma}^2_e$), whereas the numerator of $\hat{\sigma}^2_p$ is subtracted by the random-error term only in the nonadditive model. Consequently, if an additive model is inadvertently used to analyze nonadditive data, the estimated variance component for the person effect can be adversely underestimated. Zhang and Lin (2013) have shown that when the degree of person-by-rater interaction is substantial, the person effect is underestimated and thereby affects the estimation of phi-coefficients (analogous to reliability coefficients in classical test theory) in G-theory framework. They have further shown that in some cases, the underestimation can result in negative variance estimates, which is against the concept of variance component. On the other hand, when the interaction is insignificantly small or does not exist (i.e., $\sigma^2_{pr} = 0$), the additive and nonadditive models do not differ with respect to variance component estimates.

5.2 Method

As a follow-up to the study by Zhang and Lin (2013), the current study attempts to explore statistical tests that can adequately detect nonadditivity in data so that the correct use of G-theory models, be it additive or nonadditive, can be better assured. Tukey's single-degree-freedom test for nonadditivity (Tukey, 1949) is the only approach to date that is developed for the single-observation-per-cell type of measurement in testing the significance (or lack thereof) of nonadditivity in data. The logic behind Tukey's test is briefly sketched here—readers are directed to Tukey (1949) for full details on the statistical procedures. First, Tukey's test isolates the sum of squares of a single-degree-freedom nonadditive interaction contrast from the sum of squares of the confounding error component ($\varepsilon_{pr,e}$). Second, it performs a hypothesis test (i.e., H_o: $\sigma_{pr}^2 = 0$, H_1: $\sigma_{pr}^2 \neq 0$) regarding the nonadditive interaction contrast via an *F* ratio statistic:

$$F_{\text{Tukey}} = \frac{\text{SS}_{pr}/1}{(\text{SS}_{pr,e} - \text{SS}_{pr})/(df_{pr,e} - 1)},$$

where SS_{pr} is the observed sum of squares of the nonadditive interaction, $\text{SS}_{pr,e}$ is the observed sum of squares of the error component, and $df_{pr,e}$ is the degree of freedom associated with $\text{SS}_{pr,e}$. The observed *F* ratio is to be compared with $F_{.05}(1, df_{pr,e} - 1)$. A lack of significance for the interaction contrast would lend support to additivity (i.e., H_o), while a significant interaction contrast points to nonadditivity (i.e., H_1).

What have not been investigated much in the literature are the Type I and Type II error rates of Tukey's single-degree-freedom test for nonadditivity and its applications in the G-theory framework. In view of the impact of nonadditivity on G-theory analysis, the purposes of the current study are to:

- evaluate Tukey's single-degree-freedom test for nonadditivity in terms of Type I and

Type II error rates;

- demonstrate the application of Tukey's test to an operational dataset and highlight its usefulness in correcting for the underestimation of variance components when it occurs; and
- more importantly, develop a statistical program in performing Tukey's test and in making subsequent adjustments for the underestimation of variance components.

The current study evaluates Tukey's test in terms of Type I and Type II error rates via a Monte Carlo simulation study; in addition, it applies Tukey's test to an empirical judgmental study of educational standards, in which a panel of expert reviewers rated a set of standards using an established cognitive scale. The analysis in this study was performed in the R statistical software, version 2.15.2. Details on the simulation and empirical studies are provided next.

Type I error rate is defined as the chance of falsely rejecting a null hypothesis when in fact it is true (Howell, 2013). Generally speaking, for a statistical test to be considered useful, the Type I error rate should be small and is usually set at .05 as a rule of thumb. By situating the notion of Type I error rate in the current study, it translates to the probability of Tukey's test in showing erroneous significant interaction effects for nonadditivity when the data are actually additive. Given that, data generation for the purpose of evaluating the Type I error of Tukey's test followed the assumption of additivity, such that the three random effects in Model (5.1) were generated independently from three normal distributions, where $\alpha_p \sim N(0, \sigma_p^2)$, $\beta_r \sim N(0, \sigma_r^2)$, and $\varepsilon_{pr,e} \sim N(0, \sigma_e^2)$, respectively. By generating the three random components independently of one another, one can be certain that nonadditivity is not present because nonadditivity occurs only when some or all of the components are correlated.

Type II error rate is defined as the chance of failing to reject the null hypothesis given that the null hypothesis is actually false (Howell, 2013). What might be more intuitive in the discussion of Type II error rate is the notion of statistical power. Power is defined as the probability of accurately rejecting the null hypothesis when in fact it is false. It is clear from the definitions that power and Type II error rate will sum up to 1. Generally, high power for a statistical test is desirable, and satisfactory power is usually set at .80. Applying the notion of statistical power to Tukey's test would indicate its ability to accurately detect significant nonadditive interactions when the data are in fact nonadditive.

5.3 A Simulation Study

5.3.1 Simulation designs. In both the Type I error analysis and the power analysis of Tukey's test, four sample sizes (n_p): 25, 50, 100 and 1,000, and 4 numbers of raters (n_r): 3, 5, 10 and 20 were selected; therefore, a total of 16 conditions were considered in the simulation study. Myers (1979) argued that Tukey's test was particularly sensitive to "correlation between a subject's average performance and the rate at which his performance changes relative to the changes in the group performance" (p. 185). In a rated assessment, this correlation would suggest that for high-performing persons, lenient raters are likely to award higher scores while harsh raters tend to be more conservative in their scoring. For low-performing students, lenient raters would not uniformly give higher scores; likewise, harsh raters would not necessarily assign lower scores. Due to the different rating patterns in relation to how a person performs, a significant person-by-rater interaction exists and thereby constitutes nonadditivity.

Data generation for the purpose of evaluating the statistical power of Tukey's test aimed to incorporate the above correlation identified by Myers (1979). Suppose that one is working with a *p*-by-*r* data matrix, where persons constitute the rows and raters the columns. This

correlation is operationalized as that between the average of person scores across all raters ($\bar{X}_{p.}$) and the sum of cross-products of the person score and the deviation of average rater score from the overall mean ($\sum_r X_{pr}(\bar{X}_{.r} - \bar{X}_{..})$). In the current simulation study, this correlation was targeted at .50 so that it represents a medium magnitude of nonadditivity. The actual average correlation was .54 across all the simulated conditions. The correlation was realized by adding an interaction effect ($\alpha\beta_{pr} \sim N(0, \sigma^2_{pr})$ in Model (5.1) so that the interaction correlated with both the person and rater effects. By allowing the person and rater effects to be correlated with the interaction effect, the simulated data become nonadditive.

5.3.2 Results. For each simulated condition, 1,000 replications were conducted. Hence, Type I error is calculated as the number of replications out of 1,000, in which Tukey's test erroneously suggests the presence of nonadditivity, whereas power is calculated as the number of replications out of 1,000, in which Tukey's test is successful in picking up nonadditivity in the data. Table 5.2 presents results of Type I error rates for Tukey's test across the 16 simulated conditions. Results show that the Type I error rate of Tukey's test is around .05 for each condition considered in the current study, suggesting that the test has a satisfactory control for occurrences of falsely detecting nonadditivity when the data are actually additive.

Table 5.3 shows the results of power analysis for Tukey's test. As expected, when the number of persons is fixed, the power increases as the number of raters increases. For example, when $n_p = 50$, the power increases from .69 to 1.00 as the number of raters increases from 3 to 20. In a similar vein, when the number of raters is fixed, the power improves with more persons. As can be observed from the average power for each sample size (i.e., n_p), results show that the power of Tukey's test is above .80, indicating that the test is sensitive to the type of nonadditive interaction suggested by Myers (1979) when it in fact exists.

5.4 An Empirical Study

5.4.1 Empirical data. Empirical data were collected in 2009 during a judgmental study of educational standards in Oklahoma (Cook, Wilmes, Chi, & Lin, 2009). One of the objectives of the study was to rate the cognitive complexity of a set of 25 standards using the Depth of Knowledge (DOK) scale developed by Webb (2002). On a scale of 1 to 4, trained reviewers gave DOK ratings based on the content/task represented in the standards. Level 1 is the lowest level, representing low cognitive-demand processing, while level 4 indicates high-level complex processing. A panel of four reviewers ($n_r = 4$) rated each standard independently based on the established cognitive scale. The standards and reviewers were crossed in the dataset; that is, each reviewer rated the same set of 25 standards. Different from a typical rated assessment in which persons are usually the objects of measurement, the standards in this study were treated as the objects of measurement ($n_p = 25$). In the empirical study, the extent to which the panel of reviewers reliably interpreted the standards in a similar way was of primary interest. The phi-coefficient, a reliability-like coefficient in G theory, was adopted to serve this purpose and is computed as follows:

$$\text{phi} - \text{coefficient}\ (\Phi) = \frac{\hat{\sigma}_p^2}{\hat{\sigma}_p^2 + \hat{\sigma}_r^2 + \hat{\sigma}_{pr,e}^2}. \tag{5.2}$$

Recall from Table 5.1 that the additive and nonadditive models differ in the estimation of variance component for persons ($\hat{\sigma}_p^2$). As such, the underestimation of $\hat{\sigma}_p^2$, due to failing to consider nonadditivity when the data is in fact nonadditive, will also result in the underestimation of the phi-coefficient. Following the suggestion by Scheffe (1999) regarding data scrutiny, the empirical analysis first examined the relative magnitude of the variance component for errors ($\hat{\sigma}_{pr,e}^2$). Upon finding a relatively large value for the error component,

Tukey's test for nonadditivity was performed. A significant F_{Tukey} would then inform the correction for the underestimation of $\hat{\sigma}_p^2$. R syntax for the empirical analysis is appended at the end. Note that the code is written in a generic fashion such that it can be readily applied to any complete one-facet dataset so long as the dataset is arranged in an object-of-measurement-by-facet format.

5.4.2 Results. For illustrative purposes, in addition to examining the relative magnitude of $\hat{\sigma}_{pr,e}^2$, the variance components for persons and reviewers were estimated using the one-facet additive model regardless of the nature of the data. The mean squares, estimated variance components, and their respective proportions of total variance are reported in Table 5.4. First, it is obvious that the relative magnitude of $\hat{\sigma}_{pr,e}^2$ is large in that it accounts for 47.4% of total variance, which is a first hint of potential nonadditivity in the data. Second, it is odd to observe a negative $\hat{\sigma}_p^2$ because it is against the notion of variance component, which further warrants the use of Tukey's test to detect nonadditivity given that the underestimation of $\hat{\sigma}_p^2$ in the presence of nonadditivity may lead to the negative value.

Next, Tukey's test shows a significant nonadditive interaction, $F_{\text{Tukey}}(1, 71) = 221.098$, $p < .001$. It indicates that the one-facet nonadditive model should be used instead in the analysis. To obtain the $\hat{\sigma}_p^2$ under the nonadditive model (i.e., the nonadditive variance component for persons), one needs to first estimate σ_e^2 via the partial omega squared for the nonadditive interaction contrast ($\omega_{(pr)}^2$). From the definition of $\omega_{(pr)}^2$ (see, Keppel & Wickens, 2004, p.165; Zhang & Lin, 2013), it is the ratio of σ_{pr}^2 to $\sigma_{pr,e}^2$; hence, in relation to the nonadditive variance component for persons from Table 5.1, some algebraic manipulation would yield that $\hat{\sigma}_e^2$ can be replaced by $\hat{\sigma}_{pr,e}^2(1 - \hat{\omega}_{(pr)}^2)$. Next, one needs to estimate the partial omega squared based on the

observed F ratio for the interaction contrast (F_{Tukey}) as follows:

$\hat{\omega}^2_{(pr)} = (F_{\text{Tukey}} - 1)/(F_{\text{Tukey}} - 1 + 2n_p)$. The $\hat{\sigma}^2_p$ for nonadditive model is then: $\hat{\sigma}^2_p = [\text{MS}_p - \hat{\sigma}^2_{pr,e}(1 - \hat{\omega}^2_{(pr)})]/n_r$. Table 5.5 presents the mean squares, estimated variance components, and their respective proportions of total variance based on the one-facet nonadditive model.

Because of the significantly large nonadditive interaction contrast identified by Tukey's test, one can observe that the underestimation of variance component for persons based on the additive model ($\hat{\sigma}^2_p = -.005$) has now been corrected upward based on the nonadditive model ($\hat{\sigma}^2_p = .017$) by comparing Table 5.4 with Table 5.5. Next, by plugging in the estimated variance components into Equation (5.2), one can obtain the phi-coefficient to assess the reliability of the panel of reviewers in interpreting the standards. Had the additive model been used in the analysis, the phi-coefficient would have been -.022. With the correction of $\hat{\sigma}^2_p$ based on the nonadditive model, the phi-coefficient has now become .069.

5.5 Discussion

The current study seeks to advance the discussion of nonadditivity in the context of G-theory applications. It has shown analytically and empirically that when nonadditivity is present, the variance component for persons can be underestimated. In addition, it evaluates Tukey's test for identifying nonadditive data in terms of Type I and Type II error rates and demonstrates the correction for underestimation based the test results. More importantly, the study illustrates the usefulness of Tukey's test in a G-theory application and offers statistical program code for performing the relevant analysis.

Just as any statistical test, although Tukey's test is not perfect in the sense that "[it] will not be sensitive to all interactions" (Myers, 1979), it is nevertheless an effort to address

nonadditivity given the complications introduced by it. Future research can seek to investigate the test's sensitivity (or lack thereof) to various types of nonadditive interactions and to develop other procedures that can complement Tukey's test when it fails to detect nonadditivity.

CHAPTER 6

CONCLUSION AND FUTURE RESEARCH

The current dissertation looked into methodological challenges and application issues in G theory. Specifically, it deals with various existing realities related to the use of G theory as an analytical framework in rated measurement by human raters. First, the existing reality of sparse rated data in performance-based assessments has prompted researchers to come up with different analytical approaches to estimating variance components under the G-theory framework. What has not been researched much is the comparison of estimation precision based on the different approaches, which is a gap that the current dissertation aims to fill. Second, this dissertation tackles the existing reality of recruiting sufficient numbers of expert reviewers in judgmental studies of educational standards. As such, it advances new applications of G theory in the context of English language proficiency (ELP) standards-to-standards correspondence research. Finally, the current dissertation touches on the issue of nonadditivity in the G-theory framework, which is not sufficiently discussed in the G-theory literature. Nevertheless, other unsolved issues remain with respect to the use of G theory in rater-mediated measurement. This chapter serves as a discussion of some of the unsolved issues and provides ideas for future research.

6.1 Generalizability Theory and Many-Facet Rasch Measurement

Regarding the analysis of expert rated assessment of actual test performances, G theory operates at a macro-level by parsing out global information about the overall score variability among raters (i.e., estimated variance component for raters). What might be equally important in such a context, particularly for quality assurance of rater training, is to also look at individual rater variability at a micro-level. Given that, a popular measurement model widely used in the field of language testing to complement the macro-level G-theory analysis is the many-facet

Rasch measurement (MFRM) (Linacre, 1989; Wright & Master, 1982). MFRM is a branch of latent-trait modeling which relates examinee observed scores to their latent trait (i.e., underlying ability). Specifically, MFRM conceptualizes an examinee's response score as a function of his/her latent ability and a set of measurement facets. In the context of performance-based language assessments, these facets usually correspond to raters, tasks, and scoring categories. Many studies have utilized both G theory and MFRM in the analyses of performance-based language tests (e.g., Akiyama, 2001; Bachman et al., 1995; Lynch & McNamara, 1998; Sudweeks et al., 2005). G-theory analyses provide group-level information about the overall effects of various facets involved in rated measurement and contribute to decision-making about test designs, while MFRM analyses can also investigate these facets but are useful in offering more fine-grained information about specific rater or task effects that may benefit test development. Although results based on both approaches may not always be consistent with each other (Bachman et al., 1995; Lynch & McNamara, 1998), the two approaches should not be considered at odds. Rather, G-theory and MFRM analyses complement each other. As Lynch and McNamara (1998) have nicely put, "[o]ne way of reconciling these apparent differences is to recognize that the [two approaches] operate with differing levels of detail" (p. 176).

Moreover, the computer program FACETS (Linacre, 2010) has promoted the popularity of MFRM in examining individual rater/task effects in the field of language testing (see Eckes, 2011; McNamara, 1996). This commercial program has a free version for data points up to 2,000. Despite the wide application of MFRM in evaluating individual rater/task effects in performance-based language tests, a closer examination of typical assessment designs, in which multiple ratings are given to each response, reveals some insufficiencies of MFRM in this context. As such, the MFRM modeling ignores response dependence among ratings of the same

piece of examinee work. In other words, model specifications in MFRM do not fully capture potential rating dependencies in multiple-rating designs. MFRM is one variant of item response theory (IRT) modeling (Hambleton & Swaminathan, 1985), which relates a person's ability level to task characteristic(s) under a probability framework. One of the fundamental assumptions in the MFRM modeling is that multiple ratings given by raters are independent of each other given an ability level. This assumption holds in assessment designs where a single score is given to a single examinee response, which is not the typical rating procedures in performance-based assessments. However, the assumption of rating independence may not be satisfied in performance-based assessment designs in which an examinee response to a task is given multiple ratings. In essence, multiple-rating procedures in performance-based language assessments entail repeated measures of each examinee response, and hence dependencies are likely to exist among multiple ratings of the same response. By directly applying MFRM analyses in performance-based language tests with multiple-rating designs, investigators may be running the risk of not taking into account the dependence among ratings given to each response by multiple raters.

Failure to consider response dependence in multiple-rating designs has been shown to result in downward biases, i.e., underestimation, of standard errors of estimates, (Patz, Junker, Johnson, & Mariano, 2002), and also result in upward biases, i.e., overestimation, of score reliability (Wilson & Hoskens, 2001) based on the MFRM modeling. This inflation of score reliability may result in false claims about scoring consistency among raters. Given the pervasiveness of multiple-rating designs in performance-based language tests, it is important to advance the discussion of rating dependence in the field of language testing and to investigate the extent to which such dependence may affect the estimation precision of score reliability, of examinee proficiency, of rater severity, of task difficulty, and of other facets that are of interest to

language testers.

Potential alternative models would need to be able to directly account for dependencies among multiple ratings of the same examinee response so that the nature of multiple-rating procedures can be adequately specified in the alternative models. One approach to having more exact estimates in multiple-rating designs is to employ IRT-based models that specifically model response dependencies as a result of multiple-rating designs. For example, the hierarchical rater model developed by Patz et al. (2002) conceptualizes a multiple-rating design as a two-level modeling process. First, an examinee proficiency level is related to task characteristic(s) based on "ideal ratings" via a polytomous IRT model. An ideal rating is the expected score given by an ideal rater with no bias and perfect reliability. Second, observed ratings are linked to ideal ratings by a discrete signal detection process, by which "[d]ependence of ratings on various rater covariates . . . and interactions between raters and items or examinees, may also be modeled at this stage" (p. 349). Another alternative IRT-based model is the rater bundle model (Wilson & Hoskens, 2001). In this model, multiple ratings given to the same piece of examinee work are bundled together as a unit. Dependence within the unit is captured by an interaction parameter, which "is modeling a possibility that raters will agree [or disagree] more when rating this specific piece of work" (p. 288).

To my knowledge, no research that investigated individual rater/task effects in the field of language testing has considered the potential issue of dependencies among ratings given to a single examinee response in multiple-rating designs. One direction for future research is to empirically compare methods that incorporate such rating dependence with the MFRM modeling in analyzing data, in which dependencies among ratings of the same examinee work across multiple raters are clearly present. This is not to say that the MFRM modeling should be replaced

by other alternative methods. Rather, the MFRM modeling would not be the best option when the above rating dependence exists in multiple-rating designs. One way to detect such rating dependence is by using the hierarchical linear modeling (HLM) approach (see Snijders & Bosker, 2012), in which level-one units are raters and level-two units are examinees. Under the HLM approach, a score given to an examinee response can be modeled as follows:

$$X_{pr} = \mu_r + U_{0p} + R_{pr}, \tag{6.1}$$

where an examinee score (X_{pr}) is a linear combination of rater-specific intercepts (μ_r), person-specific random effects (U_{0p}), and residual errors (R_{pr}). By calculating the intra-class correlation coefficient (ICC), one can gauge the degree to which dependencies exist in multiple ratings given to the same piece of examinee response. Medium to high ICCs would suggest the use of alternative methods that incorporate such rating dependence (e.g., the hierarchical rater model and the rater bundle model). More importantly, in order to advance the discussion of rating dependencies due to multiple ratings given to each examinee work, it is imperative to develop relevant statistical programs that are publicly available so that more research efforts can continue to advance and refine research on investigating individual rater/task effects in rater-mediated measurement.

6.2 Generalizability Theory and Discrete Ordinal Scale

Given that a typical performance-based test entails examinee responses being measured on a discrete ordinal scale, one of the insufficiencies in the current G-theory framework is the lack of consideration of the discrete nature in measurement variables when one wishes to perform inferential tests that rely on distributional properties of the measurement variables. However, this does not render current G-theory applications less meaningful in any sense. The current G-theory framework is built upon ANOVA-based variance decomposition techniques. As

such, total score variability can be decomposed into estimated variance components germane to the main and interaction effects of the objects of measurement and facet(s). These estimated variance components then provide overall information about the relative magnitudes of the main and interaction effects. Based on these estimates, reliability-like coefficients are computed as measures of consistency in the measurement.

The majority of published G-theory studies that involved discrete ordinal variables focused on descriptive interpretations of the variance-component decomposition and of the reliability-like coefficients (e.g., Bachman, Lynch, & Mason, 1995; Gebril, 2009; Lee, 2006; Lee & Kantor, 2005, 2007; Lynch & McNamara, 1998; Xi, 2007). The descriptive interpretations under the current G-theory framework do not pose major methodological problems with regard to discrete ordinal measurement variables because they do not involve the distributional properties of the discrete variables; however, when one wishes to move from descriptive statistics to inferential statistics in relation to G-theory analysis, such as confidence intervals for reliability-like coefficients and hypothesis testing for variance components, the distributional properties of discrete ordinal variables need to be considered.

Due to this unsolved complication with discrete ordinal measurement scales in the G-theory framework and due to the abundance of ordinal scales in expert-rated data, investigators who work with discrete ordinal variables would need to assume that the discrete variables can be linked to some continuous latent traits, and then treat the discrete ordinal variables as continuous in performing inferential tests in relation to G-theory analyses. Nevertheless, the latent variables may be discrete or continuous, and the nature of the latent scales cannot be empirically determined.

There are alternative approaches that not only incorporate the rating dependence due to

multiple-rating designs mentioned previously but also take into consideration the discrete ordinal nature of measurement variables. An extension of the HLM approach to multiple-rating-per-examinee data in Model (6.1) is the multilevel modeling approach for ordinal measurement variables (Anderson, Kim, & Keller, 2013). One example of such an approach is the proportional odds model, also referred to as the cumulative logistic regression model, which relates probabilities of ordinal categories to predictors (e.g., rater/task effects) via a logit link function. This multilevel modeling approach for ordinal measurement scales seems promising for the development of ordinal G theory in that it considers the discrete ordinal nature of variables; however, it operates at a level of granularity different from that of the G-theory framework. As such, the multilevel modeling approach investigates rater/task effects at the individual level, whereas the G-theory framework examines rater/task effects at the group level via variance-component estimates.

Ordinal G theory has not been developed yet. One of the primary difficulties lies in the fact that the building blocks of G theory are ANOVA-based variance-component estimates, which provide group-level information about the relative magnitudes of the main and interaction effects relevant to the objects of measurement and facet(s) (e.g., raters and tasks). One direction for future research is to look for ways to jointly consider the ANOVA approach to estimating variance components and the multilevel modeling approach to analyzing discrete ordinal variables. This topic not only applies to G theory in the field of educational measurement but also is of interest to the field of statistics in general.

6.3 Generalizability Theory and Nonadditivity in Multi-Faceted Designs

In the current dissertation, it has been shown that the issue of nonadditivity (e.g., a significant person-by-facet interaction effect) affects the estimation of variance components

under the G-theory framework in a one-facet design. Nevertheless, the full potential of G theory is realized when two-facet designs or more complex ones are involved in measurement scenarios. As Lee (2006) has pointed out, the two most common facets in performance-based assessments are those associated with raters and tasks. Thus, it is of great importance to investigate ways to identify nonadditivity in multi-faceted designs and to further examine the impact of significant nonadditivity on G-theory analysis in relation to multi-faceted designs.

In the current dissertation, Tukey's single-degree-freedom test for nonadditivity is evaluated in terms of a one-facet G-theory model. As a natural follow-up research effort, future studies can extend Tukey's test in the context of two-facet designs. For instance, in a two-facet design ($p \times r \times t$), *r* refers to the facet of raters, and *t* refers to the facet of tasks:

$$X_{prt} = \mu + \alpha_p + \beta_r + \tau_t + (\alpha\beta)_{pr} + (\alpha\tau)_{pt} + (\beta\tau)_{rt} + \varepsilon_{prt,e}, \tag{6.2}$$

where the score (X_{prt}) of person *p*'s response to task *t* given by rater *r*, is the sum of an overall mean (μ), the main and interaction effects pertaining to persons (α), raters (β) and tasks (τ), and the error component (ε). It should be noted that due the single-observation-per-cell design in typical rater-mediated measurement, the three-way interaction among persons, raters and tasks, i.e., $(\alpha\beta\tau)_{prt}$, is confounded with random errors, and therefore the three-way interaction, together with the random errors, is subsumed under the error component (ε).

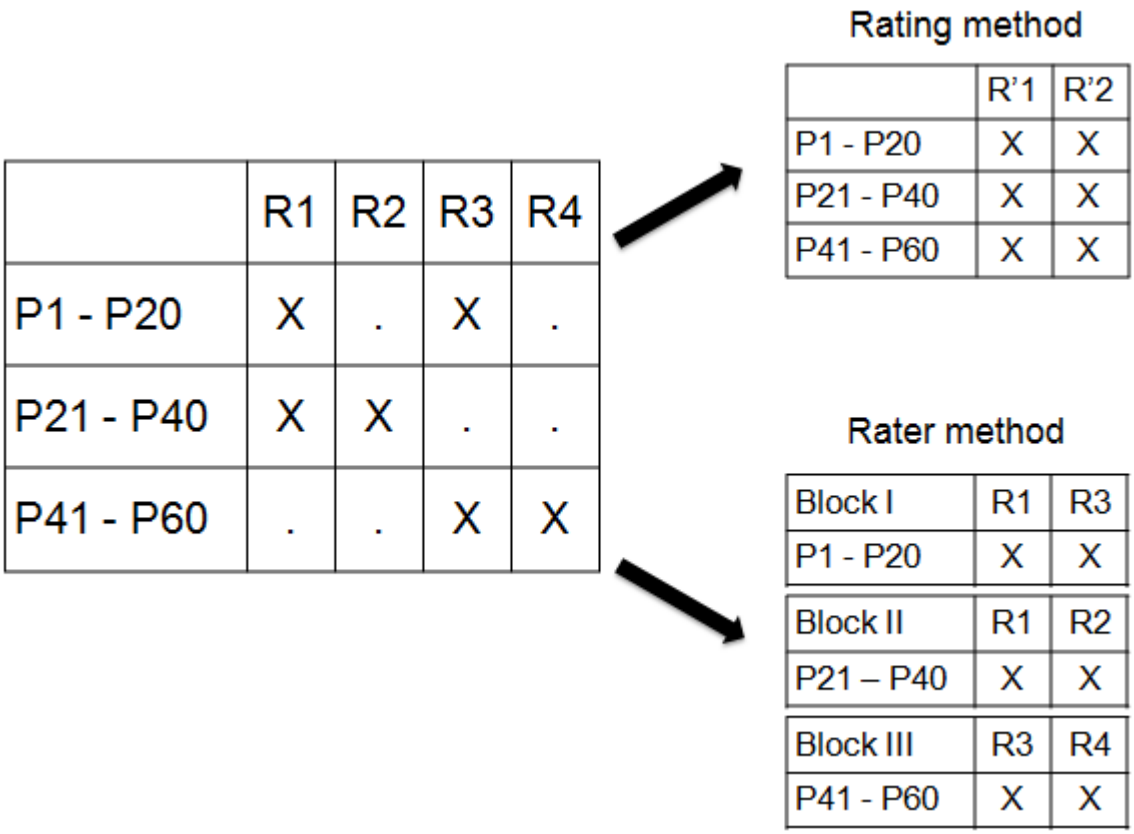
Following the logic of Tukey's test for nonadditivity, one needs to first extract the sum of squares of a nonadditive interaction contrast from the sum of squares of the confounding error component. Next, one would perform a hypothesis test (i.e., H_o: $\sigma^2_{prt} = 0$, H_1: $\sigma^2_{prt} \neq 0$) regarding the nonadditive interaction via an *F* ratio statistic:

$$F_{\text{Tukey}} = \frac{\text{SS}_{prt}/1}{(\text{SS}_{prt,e} - \text{SS}_{prt})/(df_{prt,e} - 1)},$$

where SS_{prt} is the observed sum of squares of the nonadditive interaction, $SS_{prt,e}$ is the observed sum of squares of the error component, and $df_{prt,e}$ is the degree of freedom associated with $SS_{prt,e}$. The observed *F* ratio is to be compared with $F_{.05}(1, df_{prt,e} - 1)$. A lack of significance for the interaction effect would lend support to additivity (i.e., H_o), while a significant interaction effect points to nonadditivity (i.e., H_1).

The extension of Tukey's test for nonadditivity from a one-facet design to a two-facet design appears straightforward at first glance. However, in the above two-facet example, the conceptualization of the nonadditive three-way interaction effect may not be clear, let alone the identification of the nonadditive interaction. Hence, future research would need to conceptualize nonadditive three-way interaction effects that are of importance in practice or that are known to be present in practice. In addition, one would need to find ways to extract the sum of squares of the nonadditive interaction (i.e., SS_{prt}) from the sum of squares of the error component such that SS_{prt} is independent of $(SS_{prt,e} - SS_{prt})$. The independence between the two sum-of-square terms is necessary to perform the F_{Tukey} test. In sum, exact estimates of variance components in the presence of nonadditivity require collaborative efforts from the fields of educational measurement and statistics so that psychometricians, who have the opportunity to deal with nonadditive interactions in practice, and statisticians, who have the expertise to develop statistical tests to identify such interactions in theory, could work together to further advance the use of G theory in rated measurement.

FIGURES AND TABLES



	R1	R2	R3	R4
P1 - P20	X	.	X	.
P21 - P40	X	X	.	.
P41 - P60	.	.	X	X

Rating method

	R'1	R'2
P1 - P20	X	X
P21 - P40	X	X
P41 - P60	X	X

Rater method

Block I	R1	R3
P1 - P20	X	X

Block II	R1	R2
P21 – P40	X	X

Block III	R3	R4
P41 - P60	X	X

Figure 2.1. A visual representation of the rating and rater methods

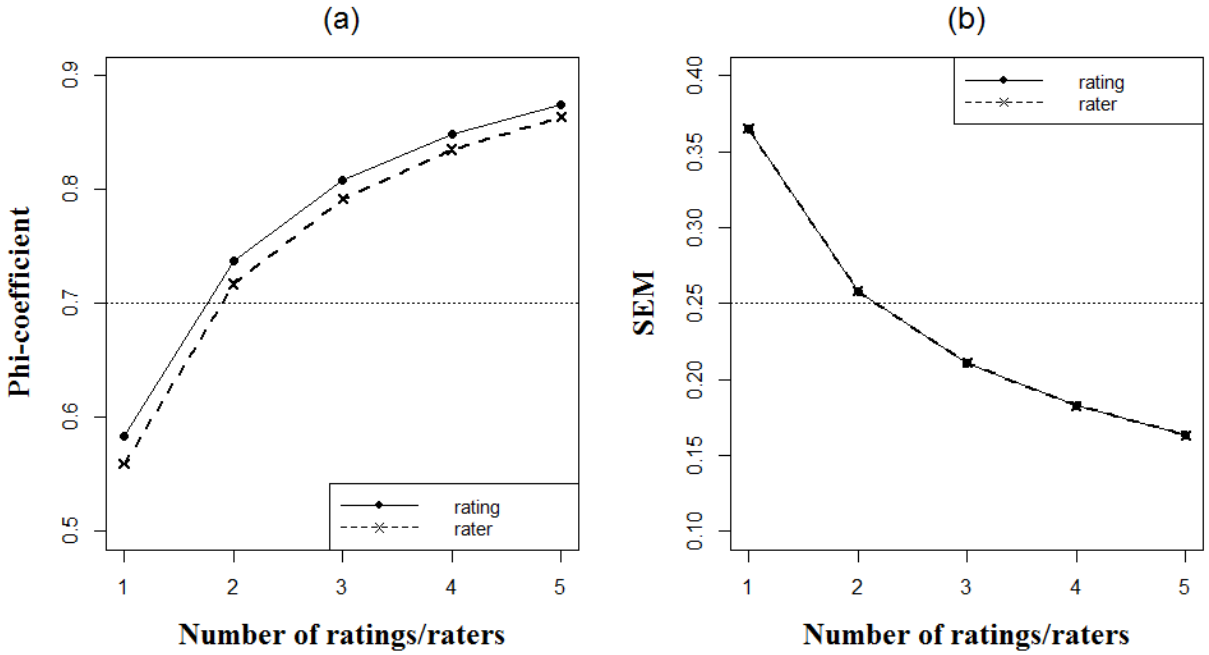


Figure 2.2. Phi-Coefficients and SEMs of EPT Writing

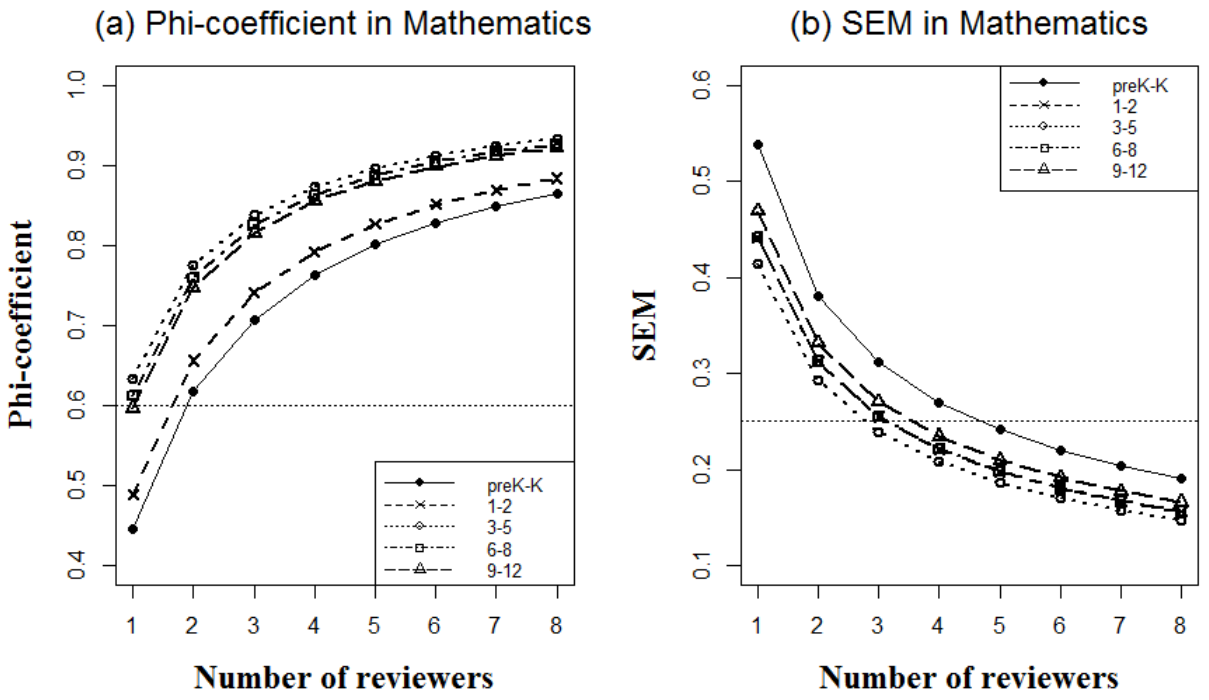


Figure 4.1. Phi-Coefficients and SEMs in Mathematics Correspondence Studies

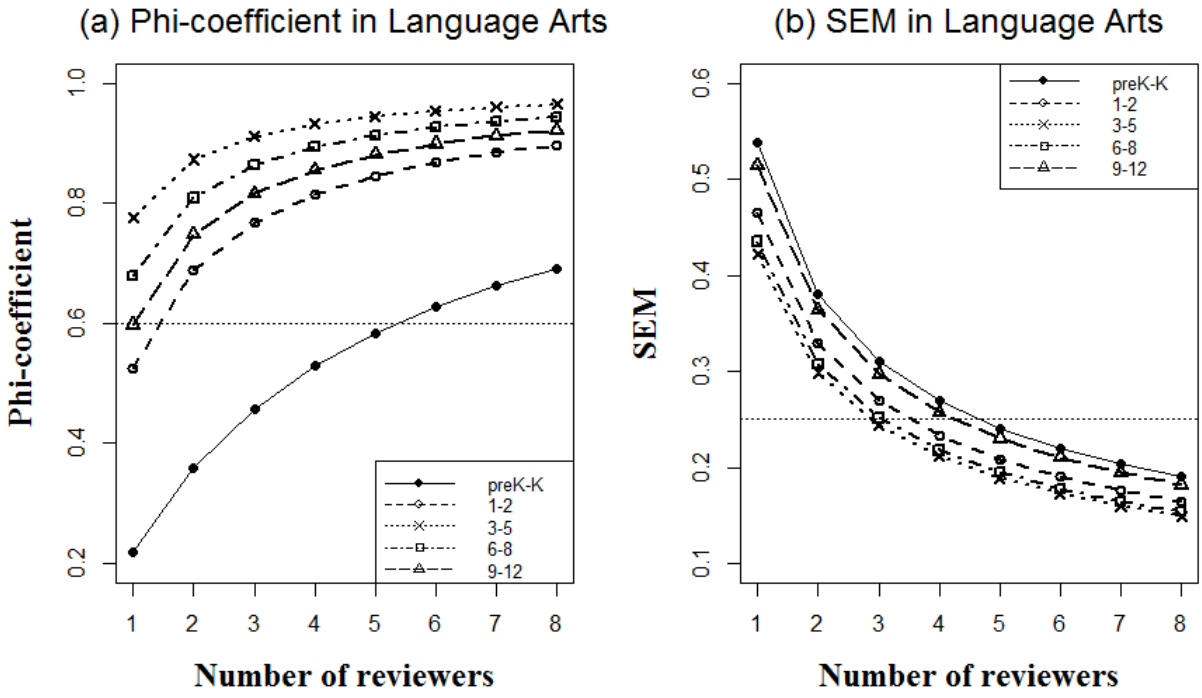


Figure 4.2. Phi-Coefficients and SEMs in Language Arts Correspondence Studies

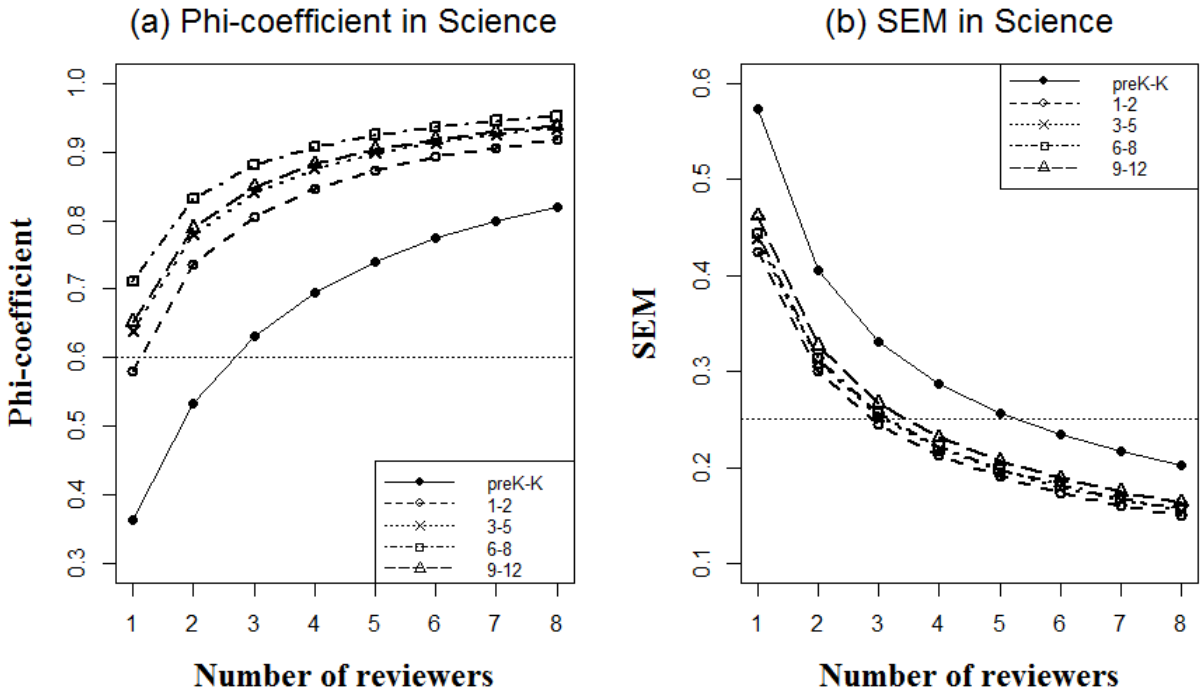


Figure 4.3. Phi-Coefficients and SEMs in Science Correspondence Studies

Table 2.1

Person Effect: Rating (Upper) vs. Rater (Lower) Methods Based on Simulations (True Parameter σ_p^2 =.4709)

n_r	$n_p = 50$			$n_p = 100$			$n_p = 200$		
	Ave. VC	**Ave. Bias**	**RMSE**	**Ave. VC**	**Ave. Bias**	**RMSE**	**Ave. VC**	**Ave. Bias**	**RMSE**
4	.4720	.0011	.1197	.4698	-.0011	.0852	.4704	-.0005	.0594
	.4720	.0011	.1210	.4697	-.0012	.0855	.4704	-.0005	.0596
8	.4716	.0007	.1213	.4712	.0003	.0836	.4706	-.0003	.0593
	.4715	.0006	.1253	.4711	.0002	.0848	.4706	-.0003	.0598
16	.4708	-.0001	.1232	.4705	-.0004	.0861	.4708	-.0001	.0587
	.4707	-.0002	.1336	.4706	-.0003	.0897	.4708	-.0001	.0597

Note. Ave. VC: average of estimated variance component over 1,000 simulations. Ave. Bias: average bias of estimated variance component over 1,000 simulations. RMSE: root mean square error of estimated variance component over 1,000 simulations.

Table 2.2

Rating/Rater Effect: Rating (Upper) vs. Rater (Lower) Methods Based on Simulations (True Parameter σ_r^2 =.0095)

n_r	$n_p = 50$			$n_p = 100$			$n_p = 200$		
	Ave. VC	**Ave. Bias**	**RMSE**	**Ave VC.**	**Ave. Bias**	**RMSE**	**Ave. VC**	**Ave. Bias**	**RMSE**
4	.00948	-.00002	.0196	.00936	-.00014	.0163	.00932	-.00018	.0153
	.00954	.00004	.0206	.00939	-.00011	.0167	.00932	-.00018	.0154
8	.00948	-.00002	.0198	.00949	-.00001	.0164	.00939	-.00011	.0145
	.00945	-.00005	.0233	.00938	-.00012	.0174	.00934	-.00016	.0148
16	.00942	-.00008	.0196	.00933	-.00017	.0165	.00963	.00013	.0152
	.00954	.00004	.0273	.00923	-.00027	.0188	.00952	.00002	.0158

Table 2.3

Error Component: Rating (Upper) vs. Rater (Lower) Methods Based on Simulations (True Parameter σ_e^2 =.2223)

n_r	$n_p = 50$			$n_p = 100$			$n_p = 200$		
	Ave. VC	**Ave. Bias**	**RMSE**	**Ave. VC**	**Ave. Bias**	**RMSE**	**Ave. VC**	**Ave. Bias**	**RMSE**
4	.2219	-.0004	.0448	.2224	.0001	.0321	.2219	-.0004	.0223
	.2218	-.0005	.0453	.2224	.0001	.0323	.2219	-.0004	.0224
8	.2220	-.0003	.0450	.2221	-.0002	.0312	.2221	-.0002	.0224
	.2220	-.0003	.0467	.2221	-.0002	.0317	.2221	-.0002	.0226
16	.2222	-.0001	.0461	.2222	-.0001	.0324	.2227	.0004	.0223
	.2220	-.0003	.0501	.2222	-.0001	.0335	.2227	.0004	.0228

Table 2.4

Person Effect: Rating (Upper) vs. Rater (Lower) Methods Based on Simulations (True Parameter σ_p^2 =.4709)

n_r	$n_p = 50$			$n_p = 100$			$n_p = 200$		
	Ave. VC	**Ave. Bias**	**RMSE**	**Ave. VC**	**Ave. Bias**	**RMSE**	**Ave. VC**	**Ave. Bias**	**RMSE**
4	.4705	-.0004	.1222	.4703	-.0006	.0852	.4714	.0005	.0605
	.4704	-.0005	.1224	.4704	-.0005	.0844	.4716	.0007	.0599
8	.4701	-.0008	.1223	.4717	.0008	.0868	.4705	-.0004	.0598
	.4704	-.0005	.1254	.4717	.0008	.0870	.4704	-.0005	.0596
16	.4705	-.0004	.1233	.4709	.0000	.0854	.4717	.0008	.0599
	.4701	-.0008	.1324	.4709	.0000	.0887	.4716	.0007	.0608

Table 2.5

Rating/Rater Effect: Rating (Upper) vs. Rater (Lower) Methods Based on Simulations

n_r	$n_p = 50$			$n_p = 100$			$n_p = 200$		
	Ave. VC	**Ave. Bias**	**RMSE**	**Ave VC.**	**Ave. Bias**	**RMSE**	**Ave. VC**	**Ave. Bias**	**RMSE**
4	.00689	-.00736	.0177	.00700	-.00725	.0153	.00694	-.00731	.0135
	.01429	.00004	.0237	.01432	.00007	.0196	.01445	.00020	.0173
	σ_r^2 =.01425								
8	.00624	-.00564	.0166	.00648	-.00540	.0136	.00654	-.00534	.0120
	.01183	-.00005	.0233	.01187	-.00001	.0172	.01185	-.00003	.0142
	σ_r^2 =.01188								
16	.00754	-.00315	.0177	.00745	-.00324	.0144	.00734	-.00335	.0125
	.01051	-.00018	.0276	.01046	-.00023	.0187	.01038	-.00031	.0144
	σ_r^2 =.01069								

Table 2.6

Error Component: Rating (Upper) vs. Rater (Lower) Methods Based on Simulations (True Parameter σ_e^2 =.2223)

n_r	$n_p = 50$			$n_p = 100$			$n_p = 200$		
	Ave. VC	**Ave. Bias**	**RMSE**	**Ave. VC**	**Ave. Bias**	**RMSE**	**Ave. VC**	**Ave. Bias**	**RMSE**
4	.2295	.0072	.0485	.2289	.0066	.0344	.2299	.0076	.0264
	.2221	-.0002	.0454	.2215	-.0008	.0315	.2224	.0001	.0222
8	.2276	.0053	.0474	.2278	.0055	.0341	.2277	.0054	.0246
	.2220	-.0003	.0472	.2224	.0001	.0328	.2224	.0001	.0226
16	.2255	.0032	.0468	.2258	.0035	.0335	.2256	.0033	.0234
	.2225	.0002	.0498	.2228	.0005	.0339	.2226	.0003	.0227

Table 2.7

Person Effect: Rating (Upper) vs. Rater (Lower) Methods Based on Simulations (True Parameter σ_p^2 =.4709)

n_r	$n_p = 50$			$n_p = 100$			$n_p = 200$		
	Ave. VC	**Ave. Bias**	**RMSE**	**Ave. VC**	**Ave. Bias**	**RMSE**	**Ave. VC**	**Ave. Bias**	**RMSE**
4	.4671	-.0038	.1202	.4641	-.0068	.0790	.4723	.0014	.0600
	.4663	-.0046	.1202	.4640	-.0069	.0786	.4717	.0008	.0584
8	.4684	-.0025	.1236	.4691	-.0018	.0855	.4674	-.0035	.0614
	.4680	-.0029	.1266	.4693	-.0016	.0857	.4676	-.0033	.0615
16	.4655	-.0054	.1233	.4714	.0005	.0871	.4703	-.0006	.0591
	.4646	-.0063	.1334	.4715	.0006	.0899	.4700	-.0009	.0596

Table 2.8

Rating/Rater Effect: Rating (Upper) vs. Rater (Lower) Methods Based on Simulations

n_r	$n_p = 50$			$n_p = 100$			$n_p = 200$		
	Ave. VC	**Ave. Bias**	**RMSE**	**Ave VC.**	**Ave. Bias**	**RMSE**	**Ave. VC**	**Ave. Bias**	**RMSE**
4	.00664	-.00760	.0173	.00744	-.00680	.0152	.00710	-.00714	.0135
	.01366	-.00058	.0226	.01481	.00057	.0204	.01410	-.00014	.0169
	σ_r^2 =.01424								
8	.01047	-.00616	.0219	.01092	-.00571	.0199	.01158	-.00505	.0192
	.01606	-.00057	.0288	.01659	-.00004	.0252	.01712	.00049	.0236
	σ_r^2 =.01663								
16	.01444	-.00337	.0271	.01499	-.00282	.0251	.01520	-.00261	.0222
	.01738	-.00043	.0359	.01813	.00032	.0295	.01848	.00067	.0256
	σ_r^2 =.01781								

Table 2.9

Error Component: Rating (Upper) vs. Rater (Lower) Methods Based on Simulations (True Parameter σ_e^2 =.2223)

n_r	$n_p = 50$			$n_p = 100$			$n_p = 200$		
	Ave. VC	**Ave. Bias**	**RMSE**	**Ave. VC**	**Ave. Bias**	**RMSE**	**Ave. VC**	**Ave. Bias**	**RMSE**
4	.2283	.0060	.0477	.2289	.0066	.0352	.2291	.0068	.0267
	.2213	-.0010	.0456	.2215	-.0008	.0312	.2221	-.0002	.0230
8	.2271	.0048	.0472	.2281	.0058	.0338	.2285	.0062	.0254
	.2215	-.0008	.0475	.2228	.0005	.0323	.2230	.0007	.0227
16	.2259	.0036	.0487	.2253	.0030	.0334	.2245	.0022	.0231
	.2230	.0007	.0512	.2216	-.0007	.0339	.2213	-.0010	.0228

Table 2.10

Descriptive Statistics of Writing Scores from EPT

	Sample size	**Mean**	**Standard deviation**	**Min.**	**Max.**
EPT	45	2.51	.55	2	4

Table 2.11

EPT Writing: Estimated Variance Components and Their Proportions of Total Score Variance (n_p=45 and n_r=4)

	***Rating* method**		***Rater* method**	
	Estimated VC	% of total variance	Estimated VC	% of total variance
$\boldsymbol{p}$	.1869	57.9%	.1692	55.9%
$\boldsymbol{r}$	-.0022*	0%	.0061	2.0%
$\boldsymbol{e}$	.1356	42.1%	.1273	42.1%
Total	.3225	100%	.3026	100%

Note. p refers to the person effect; *r* is to the rating/rater effect; *e* is the error component.
*The negative estimated variance component was carried through later calculations of phi-coefficients and SEMs as suggested by Brennan (2001).

Table 3.1

ICCs of Review Judgments about Common Core State Standards

Grade(s)	Mathematics	Grade(s)	ELA[a]-Reading	ELA-Writing	ELA-Speaking & Listening
K	0.94(0.95)	K	0.79(0.80)	0.57(0.66)	0.85(0.86)
1	0.93(0.94)	1	0.66(0.68)	0.25(0.25)	0.54(0.54)
2	0.91(0.92)	2	0.80(0.80)	0.91(0.91)	0.85(0.87)
3	0.70(0.78)	3	0.85(0.86)	0.93(0.94)	0.65(0.65)
4	0.82(0.86)	4	0.78(0.79)	0.98(0.98)	0.91(0.92)
5	0.33(0.41)	5	0.83(0.85)	0.99(0.99)	0.91(0.91)
6	0.12(0.16)	6	0.14(0.15)	0.16(0.18)	0.62(0.68)
7	0.83(0.84)	7	0.95(0.95)	0.77(0.80)	0.99(0.99)
8	0.63(0.67)	8	0.82(0.82)	0.81(0.85)	0.09(0.11)
HS-Num[b]	0.49(0.52)	9-10	0.53(0.54)	0.68(0.71)	0.62(0.62)
HS-Alg[c]	0.73(0.78)	11-12	0.73(0.77)	0.79(0.82)	0.69(0.73)
HS-Func[d]	0.55(0.57)				
HS-Geo[e]	0.66(0.68)				
HS-Stat[f]	0.60(0.66)				

[a]English Language Arts. [b]High school-Number and Quantity. [c]High school-Algebra. [d]High school-Functions. [e]High school-Geometry. [f]High school-Statistics and Probability

Table 3.2

ICCs of Review Judgments about Model Performance Indicators (MPIs)

Grade(s)	Mathematics	ELA-Reading	ELA-Writing	ELA-Speaking & Listening
K	0.91(0.92)	0.37(0.41)	0.78(0.81)	0.90(0.91)
1	0.90(0.90)	0.53(0.64)	0.64(0.67)	0.92(0.92)
2	0.92(0.93)	0.71(0.73)	0.78(0.79)	0.92(0.92)
3	0.84(0.88)	0.92(0.93)	0.95(0.95)	0.81(0.84)
4	0.82(0.88)	0.92(0.93)	0.95(0.95)	0.90(0.91)
5	0.74(0.80)	0.93(0.94)	0.95(0.95)	0.88(0.89)
6	0.90(0.90)	0.87(0.88)	0.90(0.93)	0.92(0.93)
7	0.84(0.84)	0.85(0.86)	0.89(0.92)	0.92(0.93)
8	0.82(0.83)	0.89(0.90)	0.93(0.94)	0.92(0.92)
9-10	0.81(0.82)	0.84(0.87)	0.82(0.88)	0.85(0.87)
11-12	0.81(0.82)	0.87(0.89)	0.92(0.95)	0.90(0.91)

Table 4.1

Summary of WIDA ELP Standards-to-Standards Correspondence Studies (2007-2011)

Content area	WIDA English Language Proficiency Standards	Academic content standards	Number of studies	Number of reviewers
Mathematics	The Language of Mathematics (20 MPIs)	Common Core State Standards and 10 member states' mathematics content standards	11	327
Language Arts	The Language Domain of Reading across five subareas (25 MPIs)	Common Core State Standards and 7 member states' language arts content standards	8	222
Science	The Language of Science (20 MPIs)	9 member states' science content standards	9	235

Table 4.2

Number of Random ($m \times r$) and Mixed ($m \times r \times g$) G-Study Designs

	Mathematics		Language Arts		Science	
G study	*mxr*[a]	*mxrxg*[b]	*mxr*[a]	*mxrxg*[b]	*mxr*[a]	*mxrxg*[b]
preK-K	11	0	8	0	7	0
1-2	0	11	0	8	0	7
3-5	0	11	0	8	2	7
6-8	0	11	0	8	2	7
9-12	5	6	2	6	4	5
Total	16	39	10	30	15	26

[a] *r* is a random facet. [b] *r* is a random facet and *g* is a fixed facet.

Table 4.3

Formulas for Estimated Absolute Error Variance, SEM and Phi-Coefficient

D study design	$\boldsymbol{m \times R}$ **(random)**	$\boldsymbol{m \times R \times G}$ **(mixed)**
Absolute error variance	$\hat{\sigma}^2_{\text{Abs}} = \frac{\hat{\sigma}^2_r}{n_r'} + \frac{\hat{\sigma}^2_{mr,e}}{n_r'}$	$\hat{\sigma}^2_{\text{Abs}*} = \frac{\hat{\sigma}^2_{r*}}{n_r'} + \frac{\hat{\sigma}^2_{mr,e*}}{n_r'}$
Standard error of measurement	$\text{SEM} = \sqrt{\frac{\hat{\sigma}^2_r}{n_r'} + \frac{\hat{\sigma}^2_{mr,e}}{n_r'}}$	$\text{SEM}_* = \sqrt{\frac{\hat{\sigma}^2_{r*}}{n_r'} + \frac{\hat{\sigma}^2_{mr,e*}}{n_r'}}$
Phi-coefficient	$\Phi = \frac{\hat{\sigma}^2_m}{\hat{\sigma}^2_m + \hat{\sigma}^2_{\text{Abs}}}$	$\Phi_* = \frac{\hat{\sigma}^2_{m*}}{\hat{\sigma}^2_{m*} + \hat{\sigma}^2_{\text{Abs}*}}$

** See Equation (4.1) for estimated variance components of the mixed-effect model*

Table 4.4

Estimated Variance Components in G studies for Mathematics

	preK-K (N=11)			**1-2 (N=11)**			**3-5 (N=11)**		
Source	Ave	%	SE	Ave	%	SE	Ave	%	SE
MPI (*m*)	.234	.45	.043	.186	.49	.028	.297	.63	.055
Reviewer (*r*)	.043	.08	.013	.048	.12	.020	.039	.08	.010
Error (*mr,e*)	.247	.47	.025	.147	.39	.031	.133	.29	.016
	6-8 (N=11)			**9-12 (N=11)**					
Source	Ave	%	SE	Ave	%	SE			
MPI (*m*)	.308	.61	.037	.325	.60	.034			
Reviewer (*r*)	.040	.08	.014	.032	.06	.014			
Error (*mr,e*)	.155	.31	.016	.187	.34	.029			

Note. Ave: Average of estimates across all correspondence studies. %: Proportion of estimated variance. SE: Standard error of the average.

Table 4.5

Estimated Variance Components in G studies for Language Arts

	preK-K (N=8)			**1-2 (N=8)**			**3-5 (N=8)**		
Source	Ave	%	SE	Ave	%	SE	Ave	%	SE
MPI (*m*)	.081	.22	.031	.239	.52	.035	.616	.78	.046
Reviewer (*r*)	.078	.21	.016	.063	.14	.010	.019	.02	.007
Error (*mr,e*)	.212	.57	.027	.154	.34	.016	.160	.20	.018
	6-8 (N=8)			**9-12 (N=8)**					
Source	Ave	%	SE	Ave	%	SE			
MPI (*m*)	.402	.68	.036	.393	.60	.059			
Reviewer (*r*)	.024	.04	.006	.069	.10	.022			
Error (*mr,e*)	.166	.28	.017	.196	.30	.022			

Table 4.6

Estimated Variance Components in G studies for Science

	preK-K (N=7)			**1-2 (N=7)**			**3-5 (N=9)**		
Source	Ave	%	SE	Ave	%	SE	Ave	%	SE
MPI (*m*)	.188	.37	.016	.250	.58	.039	.339	.64	.038
Reviewer (*r*)	.048	.09	.023	.038	.09	.015	.036	.07	.011
Error (*mr,e*)	.281	.54	.055	.142	.33	.029	.156	.29	.015
	6-8 (N=9)			**9-12 (N=9)**					
Source	Ave	%	SE	Ave	%	SE			
MPI (*m*)	.484	.71	.066	.340	.65	.045			
Reviewer (*r*)	.020	.03	.006	.033	.05	.012			
Error (*mr,e*)	.177	.26	.033	.181	.30	.025			

Table 4.7

Number of Reviewers Recommended Based on Three Approaches by Grade Clusters

	Phi-coef only			**SEM only**			**Phi-coef and SEM**		
	MA	*LA*	*Sci*	*MA*	*LA*	*Sci*	*MA*	*LA*	*Sci*
preK-K	2	6	3	5	5	6	5	6	6
1-2	2	2	2	4	4	3	4	4	3
3-5	1	1	1	3	3	4	3	3	4
6-8	1	1	1	4	4	4	4	4	4
9-12	2	2	1	4	5	4	4	5	4

Table 4.8

Number of Reviewers Recommended Based on Three Approaches by Grade Levels

	Phi-coef only			**SEM only**			**Phi-coef and SEM**		
	MA	*LA*	*Sci*	*MA*	*LA*	*Sci*	*MA*	*LA*	*Sci*
preK-K	2	6	3	5	5	6	5	6	6
1	3	3	2	5	5	5	5	5	5
2	3	2	2	5	5	4	5	5	4
3	2	1	2	6	5	5	6	5	5
4	2	1	2	5	4	5	5	4	5
5	2	1	2	5	4	4	5	4	4
6	2	2	1	5	5	5	5	5	5
7	2	1	1	5	5	4	5	5	4
8	2	2	1	5	5	4	5	5	4
9-12	2	2	1	4	5	4	4	5	4

Table 5.1

Estimated Variance Components for One-Facet (pxr) Additive and Nonadditive Models

	Additive model	**Nonadditive model**
Person (*p*)	$\hat{\sigma}_p^2 = \frac{\mathrm{MS}_p - \hat{\sigma}_{pr}^2 - \hat{\sigma}_e^2}{n_r}$	$\hat{\sigma}_p^2 = \frac{\mathrm{MS}_p - \hat{\sigma}_e^2}{n_r}$
Rater (*r*)	$\hat{\sigma}_r^2 = \frac{\mathrm{MS}_r - \hat{\sigma}_{pr}^2 - \hat{\sigma}_e^2}{n_p}$	$\hat{\sigma}_r^2 = \frac{\mathrm{MS}_r - \hat{\sigma}_{pr}^2 - \hat{\sigma}_e^2}{n_p}$
Error (*pr,e*)	$\hat{\sigma}_{pr,e}^2 = \mathrm{MS}_{pr,e} = \hat{\sigma}_{pr}^2 + \hat{\sigma}_e^2$	$\hat{\sigma}_{pr,e}^2 = \mathrm{MS}_{pr,e} = \hat{\sigma}_{pr}^2 + \hat{\sigma}_e^2$

Note. MS refers to observed mean squares.

Table 5.2

Type I Error Analysis of Tukey's Test Based on One-Facet Additive Model (1,000 Replications)

	n_p**=25**	n_p**=50**	n_p**=100**	n_p**=1,000**
	Type I error	*Type I error*	*Type I error*	*Type I error*
n_r**=3**	.048	.050	.045	.055
n_r**=5**	.048	.053	.054	.044
n_r**=10**	.054	.048	.042	.050
n_r**=20**	.056	.045	.050	.046
Average	.052	.049	.048	.049

Table 5.3

Power Analysis of Tukey's Test Based on One-Facet Nonadditive Model (1,000 Replications)

	n_p**=25**	n_p**=50**	n_p**=100**	n_p**=1,000**
	Power	*Power*	*Power*	*Power*
n_r**=3**	.55	.69	.75	.79
n_r**=5**	.84	.91	.94	.93
n_r**=10**	.97	.99	.99	1.00
n_r**=20**	1.00	1.00	1.00	1.00
Average	.84	.90	.92	.93

Table 5.4

Mean squares, Estimated Variance Components, and Their respective Proportions of Total Variance Based on One-Facet Additive Model

	Observed mean square	**Estimated variance component**	**Proportion of total variance**
Person (*p*)	.0892	-.005	0%
Rater (*r*)	3.13	.121	52.6%
Error (*pr,e*)	.1092	.109	47.4%

Note. The negative variance component was set to zero in the calculation of proportions

Table 5.5

Mean squares, Estimated Variance Components, and Their respective Proportions of Total Variance Based on One-Facet Nonadditive Model

	Observed mean square	**Estimated variance component**	**Proportion of total variance**
Standard (*p*)	.0892	.017	6.9%
Reviewer (*r*)	3.13	.121	49.0%
Error (*pr,e*)	.1092	.109	44.1%

REFERENCES

Adams, T. L. (2003). Reading mathematics: More than words can say. *The Reading Teachers, 56*, 786-795.

Akiyama, T. (2001). The application of G-theory and IRT in the analysis of data from speaking tests administered in a classroom context. *Melbourne Papers in Language Testing, 10*, 1-21.

Allen, M. J., & Yen, W. M. (2001). *Introduction to measurement theory*. Long Grove, IL: Waveland Press.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education (1999). *Standards for educational and psychological testing.* Washington, DC: American Educational Research Association.

Anderson, C. J., Kim, J. S., & Keller, B. (2013). Multilevel modeling of categorical response variables. In L. Rutkowski, M. von Davier, & D. Rutkowski (Eds.), *Handbook of international large-scale assessment: Background, technical issues, and methods of data analysis* (pp. 481-519). London, UK: Chapman Hall/CRC Press.

Anscombe, F. J., & Tukey, J. W. (1963). The examination and analysis of residuals. *Technometrics, 5*, 141-160.

Anstrom, K., DiCerbo, P., Butler, F., Katz, A., Millet, J., & Rivera, C. (2010). *A review of the literature on academic language: Implications for K-12 English language learners.* Arlington, VA: The George Washington University Center for Equity and Excellence in Education.

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

Bachman, L. F., Lynch, B. K., & Mason, M. (1995). Investigating variability in tasks and rater judgments in a performance test of foreign language speaking. *Language Testing, 12*, 238-257.

Bailey, A. L., & Butler, F. A. (2004). Ethical considerations in the assessment of the language and content knowledge of US school-age English learners. *Language Assessment Quarterly*, *1*, 177-193.

Bailey, A. L., & Huang, B. H. (2011). Do current English language development/proficiency standards reflect the English needed for success in school? *Language Testing*, *28*, 343-365.

Bailey, A. L., & Wolf, M. K. (2012, January). *The challenge of assessing language proficiency aligned to the Common Core State Standards and some possible solutions*. Paper presented at the Understanding Language Conference, Stanford University, CA.

Bailey, A. L., Butler, F. A., & Sato, E. (2007). Standards-to-standards linkage under Title III: Exploring common language demands in ELD and science standards. *Applied Measurement in Education, 20*, 53-78.

Bhola, D. S., Impara, J. C., & Buckendahl, C. W. (2003). Aligning tests with states' content standards: Methods and issues. *Educational Measurement: Issues and Practice, 22*(3), 21-29.

Brennan, R. L. (1992). *Elements of generalizability theory*. Iowa City, IA: The American College Testing Program.

Brennan, R. L. (2000). Performance assessments from the perspective of Generalizability Theory. *Applied Psychological Measurement, 24*, 339-353.

Brennan, R. L. (2001). *Generalizability theory.* New York, NY: Springer-Verlag.

Brennan, R. L., Gao, X., & Colton, D. A. (1995). Generalizability analyses of Work Keys listening and writing tests. *Applied Psychological Measurement, 55*, 157-176.

Brown, J. D., & Ahn, R. C. (2011). Variables that affect the dependability of L2 pragmatics tests. *Journal of Pragmatics, 43*, 198-217.

Butler, Y. G., Orr, J. E., Gutiérrez, M. B., & Hakuta, K. (2000). Inadequate conclusions from an inadequate assessment: What can SAT-9 scores tell us about the impact of Proposition 227 in California? *Bilingual Research Journal, 24*(1/2): 141–154.

Chapelle, C. A., Enright, M. K., & Jamieson, J. (Eds.) (2008). *Building a validity argument for the Test of English as a Foreign Language.* London, UK: Routledge.

Chiu, C. W. T. (2001). *Scoring performance assessments based on judgments: Generalizability theory*. Boston, MA: Kluwer Academic.

Chiu, C. W. T., & Wolfe, E. W. (2002). A method for analyzing sparse data matrices in the generalizability theory framework. *Applied Psychological Measurement, 26*, 321-338.

Common Core State Standards Initiative. (2010a). *Common Core State Standards for Mathematics*. Retrieved May 31st, 2012, from http://www.corestandards.org/assets/CCSSI_Math%20Standards.pdf

Common Core State Standards Initiative. (2010b). *Common Core State Standards for English Language Arts*. Retrieved May 31st, 2012, from http://www.corestandards.org/assets/CCSSI_ELA%20Standards.pdf

Cook, H. G. (2006). Aligning English language proficiency tests to English language learning standards. In *Aligning assessment to guide the learning of all students*, (pp.135-153). Washington, DC: Council of Chief State School Officers.

Cook, H. G. (2007). *Some thoughts on English Language Proficiency Standards to Academic Content Standards Alignment*. Working Draft. Retrieved May 20th, 2012, from http:// www.nciea.org/publications/RILS_3_GC07.pdf.

Cook, H. G., & Wilmes, C. (2007). *Alignment Between the Kentucky Core Content for Assessment and the WIDA Consortium English Language Proficiency Standards.* Madison, WI: University of Wisconsin, Wisconsin Center for Education Research.

Cook, G., Wilmes, C., Chi, Y., & Lin, C. (2009). *Alignment between the Oklahoma Priority Academic Student Skills and the WIDA Consortium English Language Proficiency Standards.* Madison, WI: Wisconsin Center for Education Research, University of Wisconsin at Madison.

Council of Chief State School Officers (CCSSO). (2002, September). *Models for alignment analysis and assistance to states.* Washington, DC. Retrieved May 1st, 2012, from http://seconline.wceruw.org/Reference/AlignmentModelsforStateAssist02.pdf

Crick, J. E., & Brennan, R. L. (1982). *GENOVA: A generalized analysis of variance system (FORTRAN IV computer program and manual).* Dorchester, MA: Computer Facilities, University of Massachusetts at Boston.

Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajaratnam, N. (1972). *The dependability of behavioral measurements*. NewYork, NY: Wiley.

Cummins, J. (1980). The construct of proficiency in bilingual education. In J. E. Alatis (Ed.), *Georgetown university round table on languages and linguistics: Current issues in bilingual education* (pp. 81-103). Washington, DC: Georgetown University.

Cizek, G. J., Bunch, M. B., & Koons, H. (2004). Setting performance standards: contemporary methods. *Educational Measurement: Issues and Practice*, *23*(4), 31-50.

Davidson, F. (2012). Test specifications and criterion referenced assessment. In G. Fulcher & F. Davidson (Eds.), *Language testing and assessment: An advanced resource book* (pp. 197-207). Oxford, UK: Routledge

Davidson, F., & Fulcher, G. (2012). Developing test specifications for language assessment. In C. Coombe, P. Davidson, B. O'Sullivan, & S. Stoynoff (Eds.), *The Cambridge guide to second language assessment* (pp. 59-65). Cambridge, UK: Cambridge University Press.

Davis, C. S. (2002). *Statistical methods for the analysis of repeated measurements*. New York, NY: Springer.

Eckes, T. (2011). *Introduction to many-facet Rasch measurement: Analyzing and evaluating rater-mediated assessments*. Frankfurt am Main, Germany: Peter Lang.

Elorbany, R., & Huang, J. (2012). Examining the impact of rater educational background on ESL writing assessment: A generalizability theory approach. *Language and Communication Quarterly, 1*, 2-24.

Ennis, R. H. (1999). Test reliability: A practical exemplification of ordinary language philosophy. *Philosophy of Education Archive*, 242-248.

Francis, D. J., Rivera, M., Lesaux, N., Kieffer, M., & Rivera, H. (2006). *Practical guidelines for the education of English language learners: Research-based recommendations for instruction and academic interventions* (No. 2). Portsmouth, NH: Center on Instruction.

Gao, X., Shavelson, R.J., & Baxter, G.P. (1994). Generalizability of Large-Scale Performance Assessments in Science: Promises and Problems. *Applied Measurement in Education, 7*, 323-342.

Gebril, A. (2009). Score generalizability of academic writing tasks: Does one test method fit all? *Language Testing, 26*, 507-531.

Gottlieb, M. (2006). *Assessing English Language Learners*. Thousand Oaks, CA: Corwin Press.

Hambleton, R., & Swaminathan, H. (1985). *Item response theory: Principles and applications*. Boston, MA: Kluwer-Nijhoff.

Herman, J. L., Webb, N. M., & Zuniga, S. A. (2007). Measurement issues in the alignment of standards and assessments. *Applied Measurement in Education, 20*, 101-126.

Howell, D. C. (2013). *Statistical methods for psychology* (8th ed.). Belmont, CA: Cengage Learning.

Huang, J. (2012). Using generalizability theory to examine the accuracy and validity of large-scale ESL writing assessment. *Assessing Writing, 17*, 123-139.

Huang, J., & Foote, C. J. (2010). Grading between the lines: What really impacts professors' holistic evaluation of ESL graduate student writing? *Language Assessment Quarterly, 7*, 219-333.

In'nami, Y., & Koizumi, R. (2013). *A meta-analysis of generalizability studies on task and rater effects in L2 speaking and writing*. Poster presented at the 35th annual Language Testing Research Colloquium (LTRC). Seoul, South Korean.

Kane, M. (1994). Validating the performance standards associated with passing scores. *Review of Educational Research, 64*, 425–261.

Kane, M. (2006). Validation. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 17–64). Westport, CT: Greenwood.

Kane, M., & Brennan, R. L. (1997). The generalizability of class means. *Review of Educational Research, 47*, 267-292.

Keppel, G., & Wickens, T. D. (2004). *Design and analysis: A researcher's handbook* (4th ed.). Upper Saddle River, NJ: Pearson Prentice Hall.

Kim, Y.-H. (2009). A G-theory analysis of rater effect in ESL speaking assessment. *Applied Linguistics, 30*, 435-440.

Kopriva, R. (2000). *Ensuring accuracy in testing for English language learners*. Washington, DC: Council of Chief State School Officers.

La Marca, P. M., Redfield, D., Winter, P. C., & Despriet, L. (2000). *State standards and state assessment systems: A guide to alignment. Series on standards and assessments.* Washington, DC: Council of Chief State School Officers.

Lee, O., LeRoy, K., Adamson, K., Maerten-Rivera, J., Thornton, C., & Lewis, S. (2008). Teachers' perspectives on a professional development intervention to improve science instruction among English language learners. *Journal of Science Teacher Education, 19*, 41-67.

Lee, Y.-W. (2006). Dependability of scores for a new ESL speaking assessment consisting of integrated and independent tasks. *Language Testing, 23*, 131-166.

Lee, Y.-W., & Kantor, R. (2005). *Dependability of new ESL writing test scores: Evaluating prototype tasks and alternative rating schemes* (TOEFL Report MS-31, RR-05-14). Retrieved from ETS® Monograph Series website: http://www.ets.org/research/policy_research_reports/rr-05-14_toefl-ms-31

Lee, Y.-W., & Kantor, R. (2007). Evaluating prototype tasks and alternative rating schemes for a new ESL writing test through G-theory. *International Journal of Testing, 7*, 353-385.

Levene, H. (1960). *Robust tests for equality of variance*. In I. Olkin (Ed.), *Contributions to probability and statistics: Essays in honor of Harold Hotelling*. Stanford, CA: Stanford University Press.

Linacre, J. M. (1989). *Many-facet Rasch measurement*. Chicago, IL: MESA Press.

Linacre, J. M. (2010). Facets Rasch measurement computer program (Version 3.67. 0) [Computer Software]. Chicago: Winsteps.com.

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Menlo Park, CA: Addison-Wesley.

Lynch, B. K., & McNamara, T. F. (1998). Using G-theory and many-facet Rasch measurement in the development of performance assessments of the ESL speaking skills of immigrants. *Language Testing, 15*, 158-180.

Martone, A., & Sireci, S. G. (2009). Evaluating alignment between curriculum, assessment, and instruction. *Review of Educational Research, 79*, 1332-1361.

Mauchly, J. W. (1940). Significance test for sphericity of n-variate normal populations. *Annals of Mathematical Statistics, 11*, 37-53.

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

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

Mooney, C. Z. (1997). *Monte carlo simulation* (No. 116). Sage.

Myers, J. L. (1979). *Fundamentals of experimental design* (3rd ed.). Boston, MA: Allyn and Bacon.

Nichols, P., Twing, J., Mueller, C. D., & O'Malley, K. (2010). Standard-setting methods as measurement processes. *Educational Measurement: Issues and Practice, 29*(1), 14-24.

No Child Left Behind Act of 2001, Pub. L. No. 107-110, 3 U.S.C. (2002).

Nugent, W. R. (2009). Construct Validity Invariance and Discrepancies in Meta-Analytic Effect Sizes Based on Different Measures A Simulation Study. *Educational and Psychological Measurement, 69*, 62-78.

Patz, R.J., Junker, B.W., Johnson, M.S., & Mariano, L.T. (2002). The hierarchical rater model for rated test items and its application to large-scale educational assessment data. *Journal of Educational and Behavioral Statistics*, 27, 341-384.

Porter, A. C. (2002). Measuring the content of instruction: Uses in research and practice. *Educational Researcher, 31*(7), 3-14.

Porter, A. C., Polikoff, M. S., Zeidner, T., & Smithson, J. (2008). The quality of content analyses of state student achievement tests and state content standards. *Educational Measurement: Issues and Practice, 27*(4), 2-14.

Porter, A. C., Smithson, J., Blank, R., & Zeidner, T. (2007). Alignment as a teacher variable. *Applied Measurement in Education, 20*, 27-51.

Rothman, R., Slattery, J.B., Vranek, J.L., Resnick, L.B. (2002). *Benchmarking and alignment of standards and testing* (CSE Tech. Rep. No. 566). Los Angeles, CA: Center for Student and Evaluation, National Center for Research on Evaluation, Standards and Student Testing (CRESST). Retrieved May 1st, 2012, from http://cse.ucla.edu/products/reports/TR566.pdf

Scheffe, H. (1999). *The analysis of variance* (Wiley Classics Library ed.). New York, NY: Wiley.

Schleppegrell, M. J. (2004). *The language of schooling*. Mahwah, NJ: Lawrence Erlbaum Associates.

Schleppegrell, M. J. (2007). The linguistic challenges of mathematics teaching and learning: A research review. *Reading and Writing Quarterly, 23*, 139–159.

Schoonen, R. (2005). Generalizability of writing scores: An application of structural equation modeling. *Language Testing, 22*, 1–30.

Shavelson, R. J., & Webb, N. M. (1991). *Generalizability theory: A primer*. Thousand Oaks, CA: Sage.

Shavelson, R.J., Baxter, G.P., & Gao, X. (1993). Sampling variability of performance assessments. *Journal of Educational Measurement, 30*, 215-232.

Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin, 86*, 420-428.

Sireci, S. G. (1998). The construct of content validity. *Social Indicators Research, 45*, 83-117.

Solórzano, R. W. (2008). High stakes testing: Issues, implications, and remedies for English language learners. *Review of Educational Research, 78*, 260-329.

Snijders, T., & Bosker, R. (2012). *Multilevel analysis: An introduction to basic and advanced multilevel modeling* (2nd ed.). London, UK: Sage.

Stemler, S. E., & Tsai, J. (2008). Best practices in interrater reliability: Three common approaches. In J. W. Osborne (Ed.), *Best practices in quantitative methods* (pp. 29-49). Los Angeles, LA: Sage.

Stoddart, T., Pinal, A., Latzke, M., & Canaday, D. (2002). Integrating inquiry science and language development for English language learners. *Journal of Research in Science Teaching, 39*, 664-687.

Sudweeks, R. R., Reeve, S., & Bradshaw, W. S. (2004). A comparison of generalizability theory and many-facet Rasch measurement in an analysis of college sophomore writing. *Assessing Writing, 9*, 239-261.

Tong, Y., & Brennan, R. L. (2007). Bootstrap estimates of standard errors in generalizability theory. *Educational and Psychological Measurement, 67*, 804-817.

Tukey, J. W. (1949). One degree of freedom for non-additivity. *Biometrics, 5*, 232-242.

U.S. Department of Education, Office of English Language Acquisition. (2003). *Final non-regulatory guidance on the Title III State Formula Grant Program – standards, assessments, and accountability*. Washington, DC: U.S. Department of Education.

Webb, N. L. (1997). *Criteria for alignment of expectations and assessments in mathematics and science education*. Council of Chief State School Officers and National Institute for Science Education Research Monograph No. 6. Madison, WI: University of Wisconsin, Wisconsin Center for Education Research.

Webb, N. L. (2002). *Alignment study in language arts, mathematics, science, and social studies of state standards and assessments for four states*. Washington, DC: Council of Chief State School Officers.

Webb, N. L. (2007). Issues related to judging the alignment of curriculum standards and assessments, *Applied Measurement in Education, 20*, 7-25.

Webb, N. L., Alt, M., Ely, R., Cormier, M., & Vesperman, B. (2005). *The Web alignment tool: Development, refinement, and dissemination*. Washington, DC: Council of Chief State School Officers.

Webb, N. M., Herman, J. L., & Webb, N. L. (2007). Alignment of mathematics state-level standards and assessments: The role of reviewer agreement. *Educational Measurement: Issues and Practice, 26*(2), 17-29.

Weigle, S. C. (1998). Using FACETS to model rater training effects. *Language Testing*, *15*, 263-287.

Weigle, S. C. (1999). Investigating rater/prompt interactions in writing assessment: Quantitative and qualitative approaches. *Assessing Writing*, *6*, 145-178.

Wilson, M., & Hoskens, M. (2001). The rater bundle model. *Journal of Educational and Behavioral Statistics,* 26, 283-306.

Wolf, M. K., Farnsworth, T., & Herman, J. L. (2008). Validity issues in assessing English language learners' language proficiency. *Educational Assessment, 13*, 80-107.

World-class Instructional Design and Assessment. (2007). *English Language Proficiency Standards*. Madison, WI. Retrieved May 1st, 2012, from http://wida.us/standards/elp.aspx#2007

Wright, B. D., & Masters, G. N. (1982). *Rating scale analysis*. Chicago, IL: MESA Press.

Xi, X. (2007). Evaluating analytic scoring for the TOEFL® Academic Speaking Test (TAST) for operational use. *Language Testing, 24*, 251-286.

Zhang, J., & Lin, C.-K. (2013). *Models in generalizability theory and negative variance components*. Manuscript submitted for publication.

APPENDIX A

Example matrix of model performance indicators (MPIs) for the Language of Mathematics in grade cluster 6-8.

	Example Topics	Level 1 Entering	Level 2 Beginning	Level 3 Developing	Level 4 Expanding	Level 5 Bridging	
LISTENING	**Percent/ Fractions**	Identify proportional representation of objects from oral directions and graphs or visuals (e.g., "Two halves make a whole. Find half a pizza.")	Follow multi-step oral directions to change proportional representation of percent or fractions in graphs or visuals	Match everyday examples of percent or fractions with oral descriptions using graphic or visual support (e.g., interest or taxes)	Analyze everyday situations involving percent or fractions from oral scenarios with graphic or visual support (e.g., "Sales tax is based on percent. When might you need to use percent?")	Apply ways of using percent or fractions in grade-level situations from oral discourse	Level 6 - Reaching
SPEAKING	**Line segments & angles**	Identify line segments or angles from pictures of everyday objects	Define or describe types of line segments or angles from pictures of everyday objects (e.g., "This angle is larger.")	Compare/contrast types of line segments from diagrams (e.g., parallel v. perpendicular lines)	Discuss how to solve problems using different types of line segments or angles from diagrams	Explain, with details, ways to solve grade-level problems using different types of line segments or angles	
READING	**Perimeter/ Area, volume & circumference**	Match vocabulary associated with perimeter or area with graphics, symbols or figures	Identify visually supported examples of use of perimeter, area, volume or circumference in real-world situations (e.g., painting a room)	Classify visually supported examples of use of perimeter, area, volume or circumference in real-world situations	Order steps for computing perimeter, area, volume or circumference in real-world situations using sequential language	Select reasons for uses of perimeter, area, volume or circumference in grade-level text	
WRITING	**Algebraic equations**	Show pictorial representations or label terms related to algebraic equations from models or visuals	Give examples and express meaning of terms related to algebraic equations from models or visuals	Describe math operations, procedures, patterns or functions involving algebraic equations from models or visuals	Produce everyday math problems involving algebraic equations and give steps in problem-solving from models or visuals	Summarize or predict information needed to solve problems involving algebraic equations	

WIDA ELP Standards (2007 edition) http://wida.us/standards/elp.aspx#2007

APPENDIX B

```
###
The following syntax is written in a generic fashion such that it can be readily applied to any
complete one-facet dataset so long as the dataset is arranged in an object-of-measurement-by-
facet format.
Note: in Chapter 5, p and r refer to persons and raters. In the following R syntax, r and c refer to
rows and columns.
###

data=read.table("X:/example/directory/data.txt")

## Obtain estimated variance component based on the one-facet additive model ##
nr=dim(data)[1]
nc=dim(data)[2]
rm=rowMeans(data)
cm=colMeans(data)
grand=mean(data)
rssq=0
for (i in 1:nr){
rssq=rssq+nc*(rm[i]-grand)^2}
cssq=0
for (j in 1:nc){
cssq=cssq+nr*(cm[j]-grand)^2}
error=matrix(0,nr,nc)
for (f in 1:nr){
for (g in 1:nc){
error[f,g]=(data[f,g]-rm[f]-cm[g]+grand)^2}}
errorssq=sum(error)
rmsq=rssq/(nr-1)
cmsq=cssq/(nc-1)
errormsq=errorssq/((nr-1)*(nc-1))
# estimated variance components #
rvar=(rmsq-errormsq)/nc
cvar=(cmsq-errormsq)/nr
errorvar=errormsq
```

```
## Tukey's test for nonadditivity ##
# Obtain the sum of squares for non-additivity #
rdev=rm-grand
cdev=cm-grand
crossproduct=c()
for (i in 1:nr){
crossproduct[i]=data[i,]%*%cdev}
Num=crossproduct%*%rdev
Dem=sum(cdev^2)*sum(rdev^2)
SSnonadd=Num^2/Dem
# Obtain F ratio#
df=(nr-1)*(nc-1)
F=SSnonadd/((errorssq-SSnonadd)/(df-1))
pvalue=1-pf(F,1,df-1)

## Obtain variance component for persons in the presence of nonadditivity ##
part_omg1=(F-1)/(F-1+2*nr)
#part_omg2=(SSnonadd/1)/(((errorssq-SSnonadd)/(df-1))+SSnonadd)
nonadd_pvar1=(rmsq-errorvar*(1-part_omg1))/nc
#nonadd_pvar2=(rmsq-errorvar*(1-part_omg2))/nc
#nonadd_pvar=(nonadd_pvar1+nonadd_pvar2)/2

## Obtain phi-coefficient ##
phi=nonadd_pvar1/(nonadd_pvar1+cvar+errorvar)
```