PROMPT AND RATER EFFECTS IN SECOND LANGUAGE WRITING PERFORMANCE ASSESSMENT

by

Gad S. Lim

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Education)
in The University of Michigan
2009

Doctoral Committee:

Professor Diane E. Larsen-Freeman, Chair
Professor Joanne F. Carlisle
Professor Nicholas C. Ellis
Jeffrey S. Johnson, English Language Institute

© Gad S. Lim

2009

ACKNOWLEDGEMENTS

Many made this dissertation and this degree come to be. I would like to thank, first of all, my committee members—Diane Larsen-Freeman, Joanne Carlisle, Nick Ellis, and Jeff Johnson—all of whom have been to me: encouraging mentors, exemplary models of scholarship, and even better human beings. I am deeply indebted to each one of you for your kindness and your support, your time and your wisdom. I owe special thanks to Diane, for being an always conscientious adviser, for always believing in me, and not least, for your uncanny knack of knowing when to use carrot and when to use stick. In addition, I would like to thank Brian Rowan, whose efforts it was that brought me to Michigan.

It was my pleasure and privilege to have worked where I have in the university, and I am grateful to the people I have worked with and learned from in each place. The Study of Instructional Improvement spirited me into classrooms and showed me the Sunnyside of school reform. To the principal investigators, to Carol Barnes and to Ruben Carriedo, and to the occupants of 3112-3119: my thanks. The English Language Institute was the birthplace of applied linguistics, and its testing division the birthplace of my own work. Besides Diane, my gratitude to Barb Dobson, Sarah Briggs, and Mary Spaan for welcoming me into the fold, and my appreciation to all those whose work—prompts written, ratings given—became the basis for this dissertation. In addition, Mark Chapman helped me code the prompts. And I cannot fail to mention the Office of the Great Seal, and in particular, those I have shared office (AO, SG, SP, SV, TK), cube (CG, EF, JC, JO, LL, RH), and broom closet (MB) with. Your friendship made it all fun, and your cupcakes made it sweet.

Thanks, too, to friends across the country and across town who kept tabs on my progress: to a dentist in the western coast, to a fellow student in Mountain time, to Thanksgivings in Central time, to a godson on the eastern coast—and to one circumambiency to the north. In Ann Arbor: to the community of Filipinos, and to the community of the Gathering, where I could park myself in more ways than one.

And then there is the place I disappeared to every summer. The Philippines. The Ateneo. Students. Friends. Family. Home. Because you provide the answer to the question why.

CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF APPENDICES

CHAPTER 1

INTRODUCTION

International examinations of English proficiency are taken by a large and ever increasing number of people every year. To name just two of the better known test programs, the International English Language Testing System, or IELTS, was administered almost a million times in 2007 (IELTS, 2008), and a similar number was estimated for the new version of the TOEFL (Cumming, Kantor, et al., 2000). Taking these examinations is often a requirement for those seeking higher education in the United States, Canada, the United Kingdom, and Australia, as well as for those seeking to migrate to these countries. At times, these proofs of proficiency in English are also important in securing employment and promotion. In other words, these tests are being used for high-stakes purposes. They not only affect the life chances of the people who take them, but also hold implications for the societies that set educational and public policy by them (McNamara & Roever, 2006; Shohamy, 2001).

In view of this, it is imperative that these exams be of the highest quality and that results obtained from them accurately represent and reflect test-takers' abilities. That is to say, these exams need to be valid, reliable, and fair. However, the reality is that there are aspects of these exams that remain imperfectly understood, where testing practice has outpaced understanding of how testing methods work. A case in point is the use of performance assessments, which is today the norm for assessing the productive language skills of speaking and writing.

Performance assessments require test takers to perform actual tasks that are similar or relevant to the knowledge, skill, or ability being measured, and success or failure on the tasks are typically judged by human raters (American Educational Research Association, American Psychological Association, and National Council on Measurement in Education [AERA, APA, & NCME] 1999; Kane, Crooks, & Cohen, 1999; McNamara, 1996). The defining characteristic is "the close similarity between the type of performance that is actually observed and the type of performance that is of interest" (Kane, et al., 1999, p. 7). The face to face oral interview, for example, is now a frequently used method for assessing speaking skills. In assessments of second language writing, it has taken the modal form of the timed, impromptu writing test (Weigle, 2002). As the name suggests, test takers have a fixed amount of time—typically thirty minutes to an hour—to write on a topic provided to them only at the time they sit for the test. In addition, Hamp-Lyons (1991) identifies five features common to such tests: (1) test takers must write at least one piece of continuous text containing no fewer than 100 words; (2) test takers are given instructions and a "prompt" which provides a general context for their writing; (3) each text is read and rated by one or more trained human judge; (4) judges' ratings are based on some common criteria, typically a set of descriptors or sample essays or one or more rating scales; and finally (5) ratings are primarily expressed as a number or set of numbers.

Performance assessments replaced discrete item and indirect tests—e.g., sentence revision tasks as a measure of writing ability—because of changes in our conception of language ability. The 1970s saw the advent of communicative approaches to language teaching (e.g., Morrow, 1979; Munby, 1978; Van Ek, 1975; Widdowson, 1978). The idea of "communicative competence" and the definition of language ability as "ability for use" (Canale, 1983; Canale & Swain, 1980; Hymes, 1967; 1972) shifted the focus away from grammatical knowledge and formal correctness towards language production and use. It followed that "the best way to test people's writing ability is to get them to write" (Hughes, 1989, p. 75). Language performance assessments, which in fact had been in use at the beginning of the twentieth century (Weir, 2003) and supplanted during the "psychometric-structuralist" era of second language testing (Spolsky 1978; 1995), were back in favor and seen as possessing greater theoretical and construct validity (Kane, et al., 1999; Linn, Baker, & Dunbar, 1991; Moss, 1992). In addition, it was argued that they had the added value of providing positive washback (Miller & Legg, 1993). That is, tests which required test takers to produce authentic language were likely to lead to similarly authentic content in language curricula and teaching.

On the other hand, there are also challenges associated with the use of performance assessments. The first has to do with the practical limits of using extended tasks. Because performance assessments tend to require more time than tests using discrete measures, examinees are typically tested on one or two tasks and scored on the basis of these limited samples. It is unclear whether performance on a small number of tasks is sufficient for representing domains as apparently complex and multi-faceted as writing and speaking ability. Thus, there is the risk of *construct underrepresentation* (Messick, 1989, 1994, 1996) in assessments of this kind, and their use raises questions about generalizability: how can one argue, on the basis of performance in one domain and one context, that the person will be able to perform at the same level in other domains and in other contexts?

Scoring is also a more difficult enterprise in performance assessment. Where scoring of discrete measures require little to no inference, scoring of performance assessments usually require the judgment of a human rater or raters. The introduction of subjectivity into the scoring process can increase *construct-irrelevant variance* (Messick, 1989, 1994, 1996), or variance due to factors not related to the construct. The traditional approach to this problem has been to calculate the statistic of inter-rater reliability. But while ways have been found to increase inter-rater reliability (Dunbar, Koretz, & Hoover, 1991), they do not necessarily make obtained scores more reliable (Henning, 1996). The desirability of increasing agreement in and of itself has come under question because agreement does not mean much if we do not know what raters are agreeing on or if they are agreeing on things unrelated to the construct (Connor-Linton, 1995a; Hake 1986; Lumley & McNamara, 1995; Reed & Cohen, 2001; Weigle, 1998). Raters of performance assessments come from many different personal and professional backgrounds; what factors they actually consider, what beliefs and predispositions they bring to the rating task, much of these remains unclear, incompletely understood, and threatens to render the ratings they give invalid.

Traditionally, validity and reliability were thought of as separate categories, where validity referred to what was being measured and reliability to how well that ability was being measured. But in the case of language performance assessment, what is being measured (e.g., writing) is also how it is being measured (i.e. through writing); language is both trait and method (Bachman, 1990). Validity and reliability are inextricably intertwined in language performance assessment, and a threat to one becomes a threat to both. Thus, one now needs to think about “scoring validity” (Chapelle, 1999; Weir, 2005). Shaw and Weir (2007) write that “scoring validity is criterial because if we cannot depend on the rating of exam scripts, it matters little that the tasks we develop are potentially valid in terms of both cognitive and contextual parameters” (p. 143).

Finally, there is the matter of providing test takers with comparable treatment. In most international exams of language proficiency, two aspects of performance assessment are systematically varied for different test takers and where test takers do not have a choice: the prompt they will respond to, and the rater who will read their responses. The one or two prompts that an individual test taker responds to are usually drawn from a larger pool of prompts for that task. It is difficult to imagine that any two prompts will be completely comparable in every way, whether in and of themselves, or in interaction with different test-taker background characteristics. Thus, how comparable are the performances of a test taker who replies to one prompt and another test taker who replies to another prompt? There is, in other words, the possibility of a “prompt effect” (Jennings, Fox, Graves & Shohamy, 1999). In the same way, there could also be a “rater effect”; test-taker responses are rated by different people, who could differ in severity and leniency, also possibly in interaction with different prompts and test-taker response and background characteristics. How comparable are scores assigned by different raters to different test takers responding to different prompts (Wigglesworth, 2007)? The issue of comparable treatment does not just raise questions about validity and reliability, perhaps more importantly, it raises questions of fairness (Kunnan, 2000). Examination providers need to make the case for the fundamental fairness, validity, and reliability of performance assessments to test takers and other stakeholders.

Fortunately, advances in theory and methodology are providing us with the framework and the tools to begin answering these questions and addressing these problems. The notion of validity itself is being elaborated and extended. Newer research and statistical methods such as verbal protocol analysis (Ericsson & Simon, 1993) and item response theory (Hambleton, Swaminathan, & Rogers, 1991) are enabling us to find out what goes on in raters' minds and to tease out the different factors that affect rater ratings.

This dissertation uses one of these newer methodologies, the multi-facet extension of the Rasch model (Linacre, 1989), in conjunction with other research methods, to explore some of the challenges brought about by the use of performance assessments in language testing. These investigations are situated in the context of one particular exam, the Michigan English Language Assessment Battery (MELAB). The MELAB is an advanced-level English proficiency test offered by the English Language Institute (ELI) of the University of Michigan to adults who use English as a second or foreign language, and is similar to the IELTS and TOEFL (ELI, 2005). Several reviews of the exam are available in the literature (Chalhoub-Deville, 2003; Purpura, 2005; Weigle, 2000). While speaking and writing are similar in that they are both productive language skills, they are also different from one another, not just on surface level features (e.g., channel, presence or absence of interlocutor, degree of co-construction), but also in social and cultural contexts of use (Brown, 1994; Grabowski, 1996; Weigle, 2002). Because what applies to one skill might not apply to the other, it is more prudent to treat writing and speaking separately. This dissertation concerns itself only with the measurement and assessment of writing ability. The main question this dissertation seeks to answer is:

> **How are the validity, reliability, and fairness of a second language writing performance assessment affected by aspects of the examination that are systematically varied for different test takers?**

The aspects that are systematically varied, as previously mentioned, are the prompts that test takers respond to, and the raters who rate these responses. In order to answer the main question, this dissertation will consider the following research questions about prompts and raters:

> Prompts: Consistent with the requirements of test validity, reliability, and fairness, to what extent are the writing prompts in a large-scale English language proficiency examination comparable in difficulty, and to what extent does the test reflect the absence of a prompt effect?

> Raters: Consistent with the requirements of test validity, reliability, and fairness, to what extent do raters in a large-scale English language proficiency examination rate appropriately and consistently, and to what extent does the test reflect the absence of a rater effect?

The rest of the dissertation constitutes an attempt to answer these questions. In Chapters 2 and 3, I consider the literature on which this study is grounded. In the former, I elaborate on the present-day understanding of validity and on how the validation of assessments involves making interpretative and validity arguments. I consider the argument for MELAB writing, and in what part of such an argument the present study is situated. The aspects of the exam that are systematically varied, as previously mentioned, are the prompts that test takers respond to, and the raters who rate these responses. In Chapter 3, I proceed to review the research on prompts and raters in relation to performance assessments to determine the frameworks that have been used, as well as to identify variables that have been investigated and variables that require investigating. The chapter ends with a presentation of the specific research questions that the study seeks to answer.

Chapter 4 provides details regarding the data and the methods used in this study. I describe the particulars of prompts, raters, and test takers in the sample of the MELAB I use. I explain the actions I take to ready the data for analysis. Following that, I explain the idea behind the Rasch model and its extension to a model that can handle multiple facets. It will be shown why this particular method is suitable for the questions the study seeks to answer. Details of the steps and procedures I follow in analyzing the data are given. The limitations of the study are also described.

In Chapter 5, I lay out the results of the study. I present and discuss findings that quantify the extent of construct-irrelevant prompt- and rater-related variance. Inferences that can be made regarding test validity will be made. Finally, Chapter 6 summarizes the findings of the study and places them within the larger language assessment research context. The chapter suggests some implications for the use and development of writing performance assessments that flow from this study, and looks forward to the farther future of writing assessment.

CHAPTER 2

VALIDATING WRITING PERFORMANCE ASSESSMENT

This chapter describes current conceptions of validity and validation in educational measurement and language testing, and applies these ideas to the validation of writing performance assessment in order to show where the current study is situated in the validation process. Two main parts make up this chapter. The first part will look at validity in general. It will be seen that validity is a theoretical notion that determines the kind of validation work that needs to be undertaken. Defining the construct to be measured is vitally important to this work. It will also be seen that doing validation work involves making validity arguments about the proposed interpretations and uses of tests, and that validation frameworks are needed to guide how evidence is collected, integrated, and evaluated. The argument structure proposed by Toulmin (1958/2003) and the frameworks developed by Kane (1992; Kane, et al., 1999) and extended by Bachman (2005) are presented as models of each. In the second part of the chapter, I consider the validation of writing performance assessment, focusing on the MELAB writing test. First, I consider the underlying construct of the test. Then, I formulate the general interpretative argument for the test and place it within a test validation framework. Finally, I home in on the section of the framework directly relevant to the present study. I look at the arguments in this section using Toulmin's argument structure, and show which parts of it the study addresses. From this, the value of the present study will be established.

Validity and Validation

Early conceptions of validity in language testing are succinctly summarized by Lado (1961) who asked: “Does a test measure what it is supposed to measure? If it does, it is valid” (p. 321). This reflected views of validity in the wider educational measurement field, which emphasized criterion validity, or the correlation between test scores and some “true” criterion measure (Cureton, 1951). The *Standards for Educational and Psychological Testing* (APA, 1954) at the time identified and discussed four types of validity: content, predictive, concurrent, and construct. The second edition of the *Standards* (AERA, et al., 1966) reduced the categories from four to three: content, criterion, and construct. These early conceptions of validity suffered from a number of weaknesses. First, the focus on the accuracy of scores came at the expense of proper consideration of the theoretical and empirical bases of these scores. Second, having different types of validity led to researchers choosing one type depending on their purposes and considering it to be sufficient for establishing validity. Finally, the purposes to which tests were employed were not examined (Brennan, 2006; Xi, 2007).

Validity

The current understanding of validity was first put forward by Messick (1989) in his seminal *Educational Measurement* chapter where he defines validity as

> an integrated evaluative judgment of the degree to which empirical evidence and theoretical rationales support the *adequacy* and

> *appropriateness* of *inferences* and *actions* based on test scores or other modes of assessment. (p.13)

This is reflected in the current edition of the *Standards* (AERA, et al., 1999), which defines validity as “the degree to which evidence and theory support the interpretations of test scores entailed by proposed uses of tests” (p. 9).

In the current conception, validity is seen as a unitary concept (“integrated evaluative judgment”) and its multiple aspects now subsumed under the banner of construct validity (Cronbach & Meehl, 1955; Messick, 1989). The construct—defined as “the concept or characteristic that a test is designed to measure” (AERA, et al., 1999, p. 5)—and construct validity are seen as key because content validity and criterion validity cannot be evaluated except by making reference to the construct. Thus, the different kinds of validity evidence are all in support of the construct and are forms of construct validity.

Validity is also not a property of a test. Rather, what are being validated are the inferences and the decisions being made on the basis of test scores. In this conception, validity does not just encompass the observations and the interpretations of them, but also the uses and consequences of tests. Messick (1989) presents these different facets under a unified view of validity in the form of a matrix (Figure 2.1).

Figure 2.1. Facets of Validity (Messick, 1989, p. 20)

	Test Interpretation	Test Use
Evidential Basis	Construct validity	Construct validity + Relevance/utility
Consequential Basis	Value implications	Social consequences

As can be seen, the evidential basis of test interpretation is construct validity. In the words of McNamara and Roever (2006), this is to answer the question: “what reasoning and empirical evidence support the claims we wish to make about candidates based on their test performance?” (p. 14). It goes without saying that first, the construct needs to be adequately defined; without it, no claims can be made. But after a construct is defined and claims are made about the relationship between observed performances and the construct, the adequacy and appropriateness of these claims need to be assessed (Cronbach, 1988). Messick (1989) identified two general threats to construct validity: construct underrepresentation and construct-irrelevant variance. In the former, observations do not include all important dimensions of the construct, whereas in the latter, observations include dimensions beyond the construct; one is a problem of measuring too little, the other a problem of measuring too much. Construct underrepresentation and construct-irrelevant variance both give rise to situations where there can be alternate interpretations of what the test is measuring.

In addition to construct validity, there are other aspects to validity. The evidential basis of test use asks if the claims and interpretations made are meaningful and appropriate given a particular context. The consequential aspects of validity concerns the social and cultural values that underlie constructs and with the societal consequences of using tests.

It follows from a definition of validity as making arguments about test interpretations and test uses that validity is always provisional, always a matter of degree, and not an all-or-nothing proposition. It involves the accumulation of evidence for particular interpretations and particular uses, but new evidence and observed consequences can support or supplant such interpretations. New contexts of use will also require new arguments about the interpretations (Fulcher & Davidson, 2009). Thus,

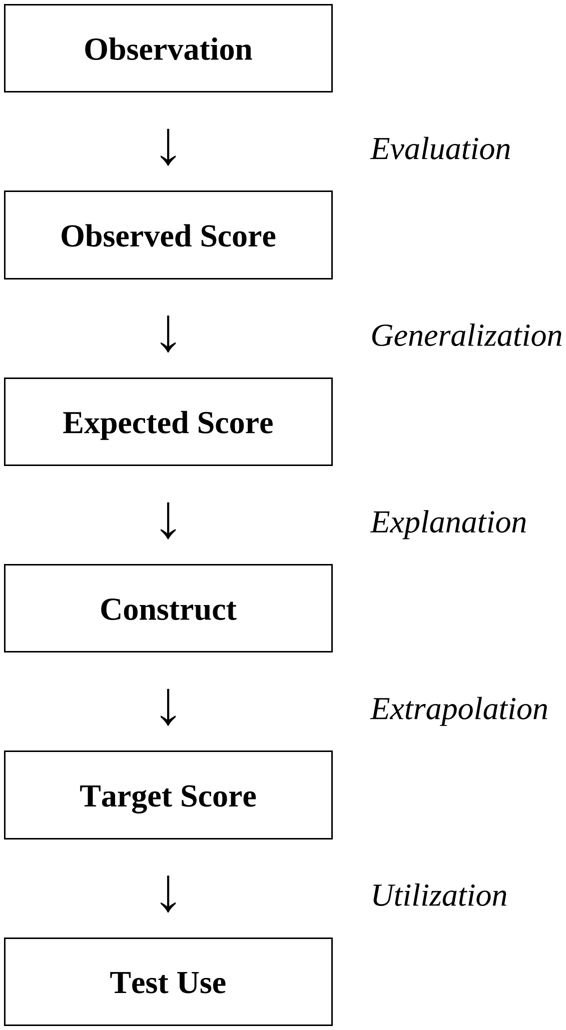
> [i]nevitably, then, validity is an evolving property and validation is a continuing process. Because evidence is always incomplete, validation is essentially a matter of making the most reasonable case to guide both current use of the test and current research to advance understanding of what test scores mean. (Messick, 1989, p. 13)

Validation Frameworks

While Messick's work has been highly influential, it has also been considered opaque and as providing little help to practitioners who need to do the work of test validation (Bachman, 2005; Brennan, 2006; Xi, 2007). It was unclear what the scope of validation work should be and how different pieces of that work might be related to each other. To the end of helping researchers and practitioners, work has been done on argument-based test validation frameworks in educational measurement by Kane (1992, 2001, 2002, 2004, 2006; Kane, et al., 1999) and extended in language testing by Bachman (2005) and by Chapelle, Enright, and Jamieson (2008), all of whom integrate Toulmin's (1958/2003) argument structure into their frameworks.

Kane (1992) proposes that an *interpretative argument* can provide a framework for collecting and evaluating evidence in support of intended score interpretations. In going from observations, i.e. test performances, to conclusions, a series of inferences are being made. Making an interpretative argument involves laying out those inferences and assumptions and evaluating them instead of taking them as givens. To illustrate (Kane, et al., 1999), the intermediate steps between an observation and a score interpretation include an observed score and a "universe" or expected score on similar tasks. Going from one component to the next involves making an inference, and the four components are linked by three inferences. This illustration of an interpretative argument, however, does not consider the place of a construct and does not account for how tests are being used. Thus, others (Bachman, 2005; Chapelle, et al., 2008) have extended this validation framework to include test constructs and test uses (Figure 2.2).

Figure 2.2 Links in an Interpretative Argument (modified after Bachman, 2005; Chapelle, et al., 2008; and Kane, et al., 1999)

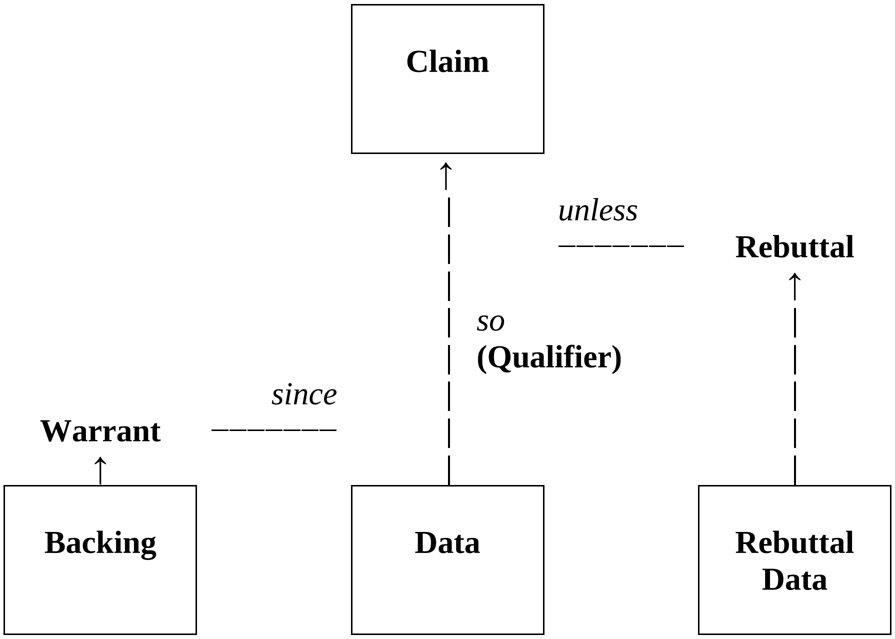


It can be seen from the figure that an interpretative argument including a test construct and test use has at least six different components which are linked together by five different inferences. The inferences allow one to move from an observation towards intermediate assertions until one arrives at an assertion about a particular test use. The first inference is *evaluation*, which is simply the scoring of the observation. This can be relatively simple and straightforward, as in marking dichotomously-scored items right or wrong according to a scoring key, or somewhat more complicated, as in the case of performance assessments where raters and rating scales are involved. *Generalization* is the drawing of an inference about an expected score for the test taker on similar tests and tasks based on the observed score. This is often thought of as reliability. The *explanation* draws a relationship between the expected score and the construct being measured. *Extrapolation* involves making an argument about a target score, or how well a test taker will perform on the construct in the real world. Finally, on the basis of a target score, decisions are made—e.g., admission into a program, placement into a course—and this is the *utilization* inference.

Bachman's (2005) version is different from the one presented here in that he does not explicitly break out the test construct and divides the whole argument into two parts, an assessment validity argument and an assessment utilization argument. The former covers test performance to the target score/score interpretation, whereas the latter covers score interpretation to score use. Other than not breaking out the explanation and extrapolation inferences, his conceptualization is no different from the one here, except to make clear and distinguish between the descriptive part and the prescriptive part of the argument. Chapelle, et al. (2008), for their part, differ from the one presented here in that they include an extra component and an extra inference. In their version, they include a target domain that comes before the collection of observations, and the inference between the two components is made by defining the domain of knowledge, skills, and attributes that the observations are supposed to reflect.

Bachman (2005), Chapelle, et al. (2008), Kane (2006), and Mislevy and his colleagues (Mislevy, Steinberg, & Almond, 2003) all propose the use of Toulmin's (1958/2003) argument structure to evaluate the overall interpretative argument as well as the specific arguments being made with each inference, in order to see how plausible they are and in order to see more clearly the potential threats to them. For Toulmin, arguments are about making claims on the basis of data and warrants (Figure 2.3).

Figure 2.3 Toulmin's Argument Structure



In Toulmin's scheme, a claim is "a conclusion whose merits we are seeking to establish" (p. 90) or the interpretation that we want to make on the basis of data. In language testing, an example of this could be about a person having a certain level of language ability. The data is any "information on which the claim is based" (p. 90); in the case of testing, this would be the test-taker's responses to the test. An inference is being made from the data to the claim on the basis of a warrant, a general statement used to justify the inference being made. A warrant is supported by backing, or "other assurances, without which the warrants themselves would possess neither authority nor currency" (p. 96). In language testing, backing can take the form of theory, research evidence, or expert judgment (Bachman, 2005). The inference can, however, be challenged by a rebuttal or conditions under which alternative explanations are possible and where the warrant would not hold. Construct underrepresentation and construct-irrelevant variance would be examples of these conditions. Rebuttal data would be evidence that helps to show the strength or weakness of alternative explanations. Finally, if desired, a qualifier can also be introduced to modulate the strength of the claim so that it is in line with the strength of the evidence presented.

Validating Writing Performance Assessment

In this section, I consider the construct of the MELAB writing test and sketch an interpretative argument for it based on the above-presented framework. I then illustrate Toulmin's argument structure and use it to situate the current research.

The Construct of MELAB Writing

There is little agreement in the field as to what exactly constitutes writing ability—and more specifically, second language writing ability. Writing appears to encompass a range of micro and macro skills, has been characterized both as a cognitive activity and as a social phenomenon, has been thought of as process and product, and has been conceived from trait, behavioralist, and interactionalist perspectives (Chapelle, 1998; Hyland, 2002; Weigle, 2002). Thus, it has been argued that "[l]ogically and empirically there seems little merit in references to a unitary construct [of] writing" (Purves, 1992, p. 112). Second language writing, whether seen from a process or product perspective, also differs from first language writing in many ways (Silva, 1993), making a definition for second language writing even more complicated an enterprise.

However many and wide-ranging the theoretical approaches to defining writing, Bachman (2007) argues that language testing practitioners work under very particular constraints. On the one hand, tests are often used for high-stakes purposes that have a large impact on the lives of many people. On the other hand, resources available for collecting evidence in support of these decisions are bounded. For these reasons, language testers "must deal with *known* constructs that may be defined very broadly" (p. 67), which coincide with the particular decisions to be made based on them, and which can be measured consistently and accurately.

The MELAB does not give a very clear statement of the construct that its writing test is measuring, which is unfortunate, but what the construct is can be inferred from various pieces of evidence. The technical manual (ELI, 2005) states that the MELAB's purpose is to "evaluate the advanced level English language competence of adult non-native speakers of English" (p. 1). A statement about its uses is also provided:

> The MELAB was developed to assess the English language proficiency of students who are applying to U.S., Canadian, British and other universities, colleges, or community colleges where the language of instruction is English. The MELAB is also used to assess the general English language proficiency of professionals such as engineers, managers, and government officials who will need to use English in their work or in on-site training. Other individuals who take the MELAB are non-native speakers interested in obtaining a general estimate of their

> English language proficiency to help them make decisions about applying for educational or employment opportunities. (p. 1)

From this statement, it would appear that the MELAB is primarily meant to measure the kind of language ability needed in educational contexts, though apparently not academic language specifically, as it is also used for professional or employment reasons. Within the context of the interpretative frameworks presented above, there are two possibilities. One is for the MELAB to define its construct more broadly in such a way that inferences based on the test can be applied to both educational and professional contexts. The other is to define academic English and professional English as separate constructs, and then develop separate interpretative arguments showing how the test is a valid measurement of each of those constructs.

Evidence more specific to the writing test is available in the writing test form (Appendix A) and in the rating scale (Appendix B). The first thing to note is that both the test form and the rating scale identify this as a test of composition, which is defined in the dictionary as "the due arrangement of words into sentences, and of sentences into periods; the art of constructing sentences and of writing prose or verse" (OED, 1996). It would appear from the MELAB's choice of words that the test is narrowing down what kind of writing it is measuring, a general ability to produce connected text. Second, it appears that productivity is one aspect of language competence; a suggested length is specified, and a note is made that extremely short compositions would be marked down. On the other hand, handwriting does not appear to be part of the writing construct; test takers are not penalized for handwriting except where their words are unreadable. This seems appropriate, as letter formation is more a concern for beginning-level writers and not for the advanced-level English that the test wants to measure. Finally, the instruction regarding outlines, that examinees may create them but that they would not be graded on them, indicates that the process aspects of writing are not included in the construct, and that the concern is with the product.

The instructions also tell the test takers what they will be marked on, namely: (1) topic development, (2) organization, and (3) range, accuracy, and appropriateness of grammar and vocabulary. This is reflected in the rating scale which, according to Weigle (2002), is “implicitly or explicitly, the theoretical basis upon which the test is founded; that is, it embodies the test (or scale) developer’s notion of what skills or abilities are being measured by the test” (p. 109). The descriptor for each scale point begins with a statement about topic development, specifying differing levels of achievement on this criterion. While topic development is being rated, topic knowledge apparently is not, as having a correct or incorrect response to prompts is not described in the scale. Thus, topic per se does not appear to be part of the construct. The statement about topic development is followed in each case by a statement on syntax and morphology, which corresponds to the “range, accuracy, and appropriateness” criterion on the test form. The descriptors for this criterion is longer for the two score points at the middle of the scale—73 and 77—suggesting that this is an important decision point and that this decision can turn on this criterion. Descriptors regarding organization, vocabulary, and spelling follow. It is unclear if there is a hierarchy to these criteria, since the order in which the criteria are presented on the test form and on the rating scale differs. In any event, the rating scale being holistic in nature should suggest that no one criterion is more important than the other, and that compositions should be evaluated as a whole.

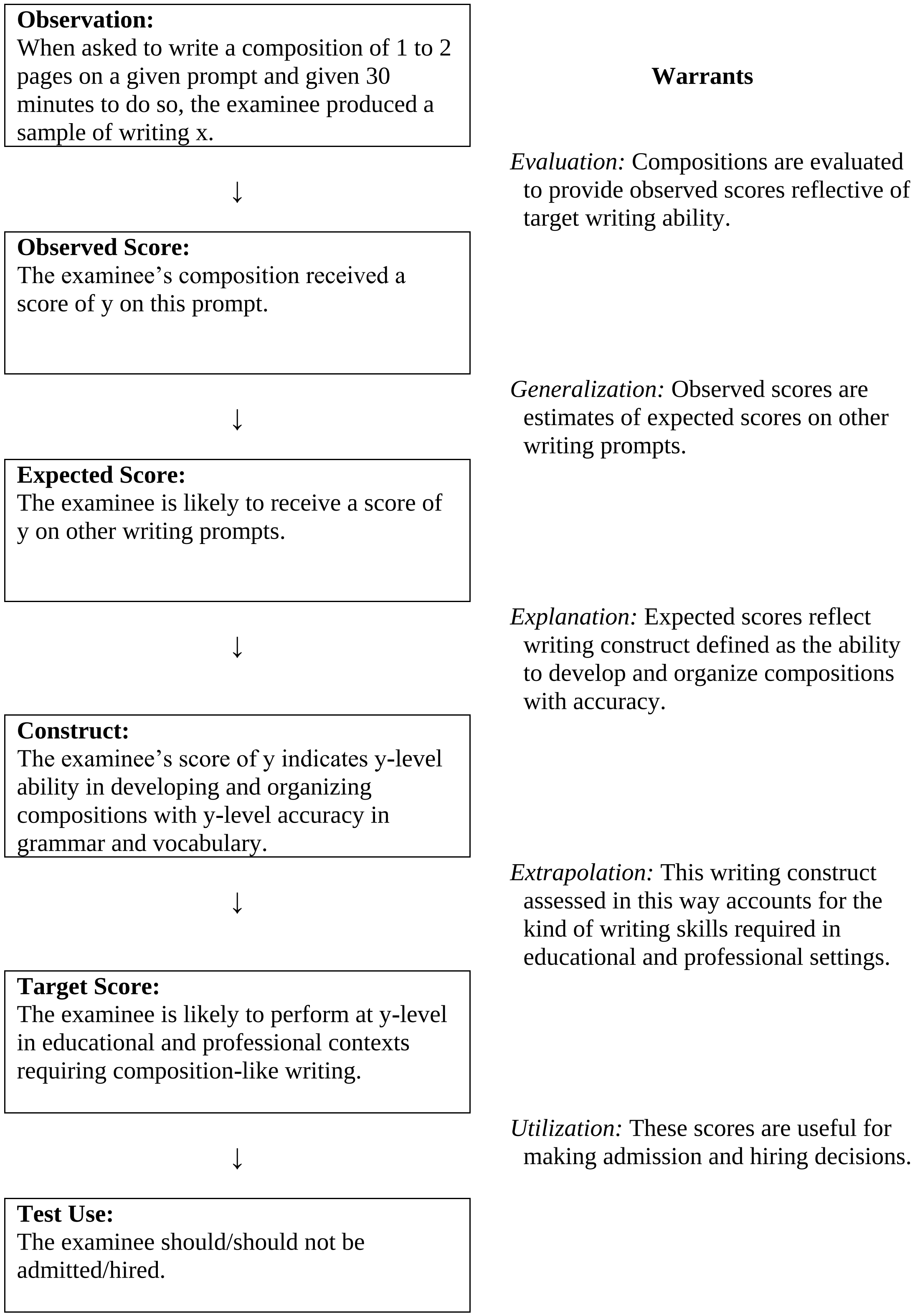
Regarding trait, behaviorist, and interactionalist approaches to construct definition (Chapelle, 1998), it would appear from the evidence that the MELAB is coming from a trait perspective. That is, consistencies in performance are attributed to test-takers' knowledge and characteristics, rather than to contextual factors, and the construct is defined in those terms. Including context as part of the construct could have been achieved through the topic, for example by instructing test takers to account for it in their responses. But as the analysis of the MELAB rating scale indicated, the topic per se does not appear to be of interest and is apparently not part of the construct. An assessment that accounted for the process aspects of writing is another avenue for specifying context and co-construction. However, the MELAB appears to be concerned with just the product of writing. From the available evidence, it would appear that the MELAB takes a trait approach to construct definition and to the nature of writing.

In sum, it would appear from the available specifications and documentary evidence that the MELAB takes a trait perspective to construct definition, and defines its writing construct to be the ability to produce a composition of some length that is appropriately organized and developed and that evidences control over different aspects of the English language. This narrowed down writing construct excludes a number of possible types and genres of writing, but is also sufficiently broad as to cover a type of writing often seen and employed in educational and professional contexts. Whether this one type of writing is sufficient to extrapolate to educational and professional writing in general probably still needs to be argued.

An Interpretative Argument for Writing Performance Assessment

Having determined what the MELAB writing construct might be, the chapter proceeds to an outline of an interpretative argument for a writing test reflecting that construct (Figure 2.4). Following the observation, the boxes on the left show the inferences being made, and to the right are the warrants which provide the reasons why the inferences are being made. Each warrant is of course based on certain assumptions, and relevant backing needs to be provided to show that the warrants are in fact warranted. Thus, for example, the warrant provided for generalizing from the observed score to the expected score is that "observed scores are estimates of expected scores on other writing prompts." This warrant is based on the assumption that prompts are created such that they are parallel or equatable to each other in some way. To provide backing for this assumption, then, reliability and equating studies need to be conducted. To give one more example of the assumptions behind warrants and the backing required for these assumptions: to extrapolate from the construct to the target score, the warrant provided is that "the writing construct assessed in this way accounts for the kind of writing skills required in educational and professional settings." The assumption underlying this warrant is that test performance is related to performance in educational and professional settings. Backing for this assumption can be provided by criterion-related validity studies, for example investigating the correlation between scores on this test and grades students receive in writing courses.

Figure 2.4 An Interpretative Argument for a Composition Writing Test (cf. Chapelle, Enright, & Jamieson, 2008)



Observation:
When asked to write a composition of 1 to 2 pages on a given prompt and given 30 minutes to do so, the examinee produced a sample of writing x.

↓

Observed Score:
The examinee's composition received a score of y on this prompt.

↓

Expected Score:
The examinee is likely to receive a score of y on other writing prompts.

↓

Construct:
The examinee's score of y indicates y-level ability in developing and organizing compositions with y-level accuracy in grammar and vocabulary.

↓

Target Score:
The examinee is likely to perform at y-level in educational and professional contexts requiring composition-like writing.

↓

Test Use:
The examinee should/should not be admitted/hired.

Warrants

Evaluation: Compositions are evaluated to provide observed scores reflective of target writing ability.

Generalization: Observed scores are estimates of expected scores on other writing prompts.

Explanation: Expected scores reflect writing construct defined as the ability to develop and organize compositions with accuracy.

Extrapolation: This writing construct assessed in this way accounts for the kind of writing skills required in educational and professional settings.

Utilization: These scores are useful for making admission and hiring decisions.

In addition to showing that the skills captured by the test are skills required in educational and professional contexts, it also needs to be shown that the skills thus captured actually represent an adequate subset or sampling of all the skills that are required in those contexts. That is to say, it is possible that the construct is represented, but that it is underrepresented. The presence of construct underrepresentation would limit the scope of the target score, and thereby affect the utility of the exam in making decisions about admission and hiring. In writing tests such as the MELAB, where only one type of writing is sampled, the question could be raised if the construct adequately covers writing in educational and professional contexts, or if the construct should perhaps be widened in some way (Shavelson, Baxter, & Gao, 1993; Weir, 2005).

The overall interpretative argument and its parts can also be framed according to Toulmin's (1958/2003) argument structure. Below, the general argument is illustrated:

- Data: the examinee's writing exhibited certain features x.
- Warrant: those features x are characteristic of y-level writing ability.
- Backing: experts who developed the rating scale judged x to be indicative of y.
- Rebuttal: unless the prompt is too difficult; unless the examinee was sick that day; etc.
- Rebuttal Data: investigation of prompt in equating study; letter from the examinee's doctor; etc.
- So + qualifier: it's highly likely that
- Claim: the examinee at y-level can/cannot write at the higher education/professional level and should/should not be accepted/hired.

In the argument structure proposed by Toulmin, all the relevant assumptions and evidence are laid out, making it easier to see the strengths and weaknesses of the argument. From the illustration, it can be seen that warrants and rebuttals are very important in the structure of an argument. Warrants spell out what inferences are being made, and help to make sure that one does not make a leap of thought that is not justified. In the same way, there are any number of rebuttals that could render a claim invalid. Those rebuttals could be of features general to the test, such as a particular prompt being more difficult than others, and could also be particular to individuals, such as in the case of an examinee being sick on the day of the test. As many rebuttals as can be investigated and alternative interpretations falsified, the stronger the claim that can be made based on the data and its warrants.

The Evaluative Inference in Writing Performance Assessment

This study is situated in the first step of the interpretative argument, where observations are given observed scores. That is, it is concerned with scores and the scoring process, or scoring validity. This step is especially crucial in performance assessments. Unlike with discrete items that can be scored using a key, which requires no inference, and where manual errors are easy to detect, the scoring of performance assessments usually requires the use of human raters who use a rating scale and exercise judgment in the awarding of marks or ratings. As well, performance assessments are often based on one or a small number of samples, making it all the more important that any scores awarded be accurate. The scoring of performance assessment is considered the weakest link in an interpretative argument, and thus requires the most attention in validation (Kane, et al., 1999; Lane & Stone, 2006).

Focusing on this part of the interpretative argument now, and using Toulmin’s (1958/2003) argument structure, the data is the composition written by the examinee, and the desired claim is that the composition should be awarded a particular score. The warrant here is that different scores are reflective of different levels of writing ability. The assumption behind this warrant is that the rating scale used to score compositions is appropriate to the task. The backing for this warrant is supplied by the judgment of the experts who developed the rating scale and deemed it an appropriate way of judging the quality of compositions. The warrant is also backed by prototyping studies conducted during rating scale development (Johnson, personal communication). Another warrant for the claim would be that the rating scale is being applied accurately and consistently.

While the backing can help to establish the warrant, rebuttals can be made, specifying scenarios and conditions under which the claim would not be justified or under which alternate explanations for the observed score are possible. It is possible to argue, for example, that a person received a score of 77 but actually deserved a score of 83, and the score of 77 was observed only because out of all the prompts available, this person only responded to one prompt and that prompt was more difficult to respond to than the others. Thus, rebuttal data needs to be presented showing that prompts are in fact comparable in difficulty or that differential difficulty is being accounted for through equating of some sort. In the same way, it can be argued that the rating scale is not being applied accurately or consistently; an observed score was higher than is deserved because the rater is not using the rating scale in the same way as other raters, and is always more lenient. Or, that the rater is more lenient on certain occasions and more severe on others. Or, that the rater is consistently more lenient towards certain prompts or test takers, but more severe towards certain other prompts and test takers. Rebuttal data are needed for these and other scenarios like it, so that strong claims about the observed score can be made based on the data.

These prompt and rater effects are examples of construct-irrelevant variance (Messick, 1989) that constitute rebuttals to the claim that examinee composition x deserves a score of y. It is vital that these rebuttals be addressed, especially as the evaluative inference is the first link in the chain of inferences, and none of the other inferences would matter if this one does not hold. As quoted in the previous chapter, "scoring validity is criterial because if we cannot depend on the rating of exam scripts it matters little that the tasks we develop are potentially valid in terms of both cognitive and contextual parameters" (Shaw & Weir, 2007, p. 143). For this reason, prompt and rater effects require looking into, and these are what the present study will investigate.

Chapter Summary

This chapter showed that validation work requires making interpretative and validity arguments, and that validation frameworks can help researchers and practitioners in this task. The construct of the writing section of the MELAB was inferred from available evidence, and the general outline of an interpretative argument for writing performance assessment was presented. The evaluative inference lies at the beginning of such an argument, and challenges to scoring validity need to be investigated. Prompt and rater effects are two important such challenges and are the focus of the present study. In view of that, the next chapter considers the literature on prompts and raters to determine the specific aspects of these that require investigation.

CHAPTER 3

LITERATURE REVIEW

This chapter examines the literature regarding prompts and raters, in particular, the relationships that different aspects of these factors might have with test-takers' outcomes in writing assessment, and the possible threats to validity that these might pose. The chapter has two parts. In the first part, the prompt factor is considered, and this consideration will begin by looking at the ways in which task difficulty has been conceived, and the reasons why predicting task difficulty has not just been difficult but also often wrong. It will be argued that difficulty should not be seen as a property of the task itself, but rather resides in the interaction between test takers and task characteristics—in this case, prompts. Following that, prompt-inherent characteristics—as opposed to difficulty features—are identified, as well as test-taker characteristics with which they might interact. A related issue that is also discussed is the provision of choice in writing assessment. The MELAB writing test is set up so that test takers are given a choice of two prompts to write about, but they write on only one. The validity implications and measurement challenges introduced by this setup are discussed. The second part of the chapter delves into research regarding raters of writing. First, it will describe a number of primarily qualitative studies on raters' rating styles and decision making processes, and models of the activity of rating that have been constructed on the basis of these studies. Then, a review is made of research into the influence of different rater qualities, characteristics, and backgrounds on rating behavior. Finally, interactions between raters and prompts are considered, including the possibility of raters compensating in their ratings for prompts they perceive to be more challenging. Having considered the literature, the chapter concludes with the specific research questions that the present study will attempt to answer.

Prompts as a Factor

As in all language use, responding to prompts requires topic knowledge. In terms of defining the assessment construct, Bachman and Palmer (1996) saw three possible ways of dealing with that fact: to define the construct solely in terms of language ability, to define it as including both language ability and topic knowledge, or to define the two as separate constructs. Depending on the way the construct is defined, topic knowledge and the variance it generates can be seen as providing information about ability or as a source of measurement error (Barkaoui, 2007a). Of the three possible definitions, the latter two are seen as most suited for assessment in language for specific purposes and vocational training programs; for language proficiency examinations such as the MELAB, where test-takers' knowledge of topics can vary widely, it is more appropriate to consider language ability alone in the construct. For tests using this definition, topic-related variance is seen as measurement error, and three suggestions have been offered for minimizing it: to use topics no test taker is expected to know, to use topics all test takers are expected to know, and to include multiple topics and either have test takers choose from the tasks or complete all of the tasks (Bachman & Palmer, 1996).

The MELAB writing test apparently employs topics that all test takers will be able to respond to according to their own knowledge, opinion, or perspective, and also includes the mechanism of choice; test takers are presented with two prompts, but only need to write on one of them. Each of these provisions is potentially problematic, however. First, regarding prompts, the relative ease or difficulty of prompts is not easy to determine. Second, regarding choice, allowing for it introduces a range of definition, performance, and measurement issues. Each of these is discussed in turn.

Prompt Difficulty

Plenty of advice has been offered with regard to the construction of performance writing assessments (e.g., Kroll & Reid, 1994; Ruth & Murphy, 1988). In spite of this, however, it seems that what makes a prompt easy or difficult still eludes people, test takers and test makers alike. For example, Freedman (1983) found that students performed no differently on writing prompts they found dull and difficult than on prompts they found easy and interesting. In another study, students allowed to choose among four prompts were shown to prefer shorter questions, but this did not result in better performance (Chiste & O’Shea, 1988). Experts are apparently not any better at making predictions. Those in a study by Hamp-Lyons and Mathias (1994) generally agreed on which kinds of prompts were easier and which ones more difficult. In agreement with the literature, it was thought that prompts calling for publicly-oriented argumentative writing would be harder than those which elicited private expository writing. The results showed the exact opposite of what was hypothesized; compositions written in response to the former category received much higher scores than those in the latter category, with combination prompts (i.e. public/expository and private/argumentative) falling in between. Robinson (1995) set out to test Long's (1985) proposal that "tasks requiring present tense, context-supported reference are simpler than those requiring the management of reference to objects and events dislocated in time and space" (p. 102). Developing a set of tasks that elicited Here-and-Now and There-and-Then narratives, Robinson tested a number of hypotheses regarding fluency, accuracy, lexis, and complexity of propositions and production. With the exception of greater lexical content for There-and-Then narratives, none of the hypotheses were supported. Other studies confirm the difficulty in judging task difficulty vis-à-vis performance outcomes (Dobson, Spaan, & Yamashiro, 2003; Greenberg, 1981; Mohan & Lo, 1985; Powers & Fowles, 1998).

The most completely conceptualized and operationalized theoretical rationale for what makes tasks easier or more difficult is that from the group working out of the University of Hawaii at Manoa (Brown, Hudson, Norris, & Bonk, 2002; Norris, Brown, Hudson, & Yoshioka, 1998). Based on the work of Skehan (1996; 1998), they have developed a matrix for evaluating and classifying the difficulty of language performance tasks across skill areas (Table 3.1). These researchers do use the notion of "task" in a different way; while most others see tasks simply as vehicles for getting at underlying language abilities, the Hawaii group subscribe to the "strong" sense of performance assessment (McNamara, 1996) where the task itself is the construct of interest, and the desired inferences are about "students' abilities to accomplish particular tasks or task types" (Brown, et al., 2002, p. 15). In any event, the different definition of task does not diminish the light shed by their taxonomy on the different dimensions of tasks that could contribute to difficulty.

Table 3.1 Language Performance Task Difficulty Matrix (Norris, et al. 1998, p. 77)

	easy → difficult	easy → difficult	easy → difficult
	range	*no. of input sources*	*delivery of input*
code	− +	− +	− +
	amount info. to process	*input/output organization*	*availability of input*
cognitive complexity	− +	− +	− +
	mode	*channel*	*response level*
communicative demand	− +	− +	− +

In the matrix, on the left-most column are the three main components contributing to task difficulty: (1) code, which refers to grammatical, textual, and pragmatic elements, (2) cognitive complexity, or the mental processes required for completing the task, and (3) communicative demand, the amount of stress involved in performing the language task. Where writing assessments such as the MELAB are concerned, many of the sub-components are clearly not relevant, as they are held constant by testing programs, e.g., *mode* is always to produce a composition, the *channel* is always writing by paper and pen, and *response level* is the same because all examinees are given the same amount of time. (The possibility of individual-level interaction with these components is certainly not discounted). The sub-components that are potentially relevant include *range*, or "the extent to which the code that is inherent in the language of a given task represents a greater or lesser degree of spread" (Norris, et al., 1998, p. 79), and amount of information to be processed and input/output organization under cognitive complexity. The caveat "potentially" relevant is required. In the MELAB, writing prompts range in length from one to five sentences. An examination of task classifications in Norris, et al. indicates that it takes gross differences for tasks to be classified as being plus or minus in difficulty in any sub-component within their scheme. Thus, prompts that differ in length by four sentences would not be classified as different under the sub-component of "amount of information to be processed" in their scheme. For the present, it might be better to err on the side of scholarly skepticism and view the four-sentence difference as being significant. On the other hand, their classifications, based as they already are on gross differences, did not prove to be particularly accurate (Brown, et al., 2002).

The problem with the above approach to difficulty, Bachman (2002, 2007) argues, is in the attempt to identify "difficulty features" in test tasks and in the attempt to conceptualize difficulty as a property of the tasks themselves. In Norris, et al., task difficulty features are defined as a combination and integration of ability requirements and task characteristics, but this appears to confound test-taker ability and test task. Of the three main components in the matrix, only code complexity is properly a characteristic or feature of test tasks. Cognitive complexity and communicative stress both require making assumptions about the test taker, and as such are properly interactions between some aspect of test task and test taker. Likewise, difficulty should not be seen as a property of the task itself. Bachman uses the athletic event of the high jump as an analogy. A bar set at five feet ten might be difficult for a high school athlete, but very easy for world-class jumpers. While the height of the bar might represent some hypothetical level of difficulty, which may be the case on average, the difficulty of the jump resides in the interaction between the height of the bar and the individual doing the jump. Difficulty is thus not a separate factor but the result of interactions among different components of assessment. The Hawaii group seems to recognize this as well. They write that the matrix is not meant to represent the universe of factors that contribute to task difficulty, and that “the different characteristics of a language event will themselves render a given task more or less difficult according to a variety of possible parameters” (Norris, et al., 1998, p. 72) – interactions, in other words.

Table 3.2 Dimensions of Tasks for Direct Writing Assessment (Weigle, 2002, p. 63)

Dimension	Example
Subject matter	self, family, school, technology, etc.
Stimulus	text, multiple texts, graph, table
Genre	essay, letter, informal note, advertisement
Rhetorical task	narration, description, exposition, argument
Pattern of exposition	process, comparison/contrast, cause/effect, classification, definition
Cognitive demands	reproduce facts/ideas, organize/reorganize information, apply/analyze/synthesize/evaluate
Specification of:	
- audience	-self, teacher, classmates, general public
- role	-self/detached observer, other/assumed persona
- tone, style	-formal/informal
Length	less than ½ page, ½ to 1 page, 2-5 pages
Time allowed	less than 30 minutes, 30-59 minutes, 1-2 hours
Prompt wording	question vs. statement, implicit vs. explicit, amount of context provided
Choice of prompts	choice vs. no choice
Transcription mode	handwritten vs. word-processed
Scoring criteria	primarily content and organization; primarily linguistic accuracy; unspecified

In view of these arguments, the more appropriate way of proceeding would be to identify not “difficulty features,” but rather prompt-inherent features (Bachman 1990; 2002)—those characteristics of prompts which require no inferences to identify and which require no assumptions about test-taker ability—and to investigate the interaction of these with various other components involved in the assessment process. One such listing of features is found in Weigle (2002) and reproduced in Table 3.2. With the exception of cognitive demands, which appears to require inferences about test-taker ability, all dimensions appear to be about features inherent in the task. The dimensions relevant to this study—that is to say, not held constant for different test takers—include subject matter, rhetorical task, grammatical person, pattern of exposition, prompt wording, and choice of prompts. These are generally the same categories that other reviews have focused on both in first language (Huot, 1990) and second language (Tedick & Mathison, 1995) writing. Each of these dimensions is discussed in turn, followed afterwards by a separate discussion of test-taker characteristics with which they can interact.

Subject Matter

The approach taken by the MELAB, as previously mentioned, is to use topics that are presumed to be familiar to all test takers. Whether this is indeed the case is worth considering. Further, even if it is the case that all test takers can write about the topic, it is very well possible that some test takers might have more expertise in a particular subject (e.g., medical professionals asked to talk about doctors) and thus have an advantage over other test takers. In Polio and Glew's (1996) study on how students choose writing topics, the most often-cited reason was having background knowledge and perceived familiarity with the topic. Almost 85% of the students gave that reason, with generality or specificity of the prompt following at approximately 46%. Background knowledge and clarity of prompt were similarly cited as reasons for choosing in Powers and Fowles (1998).

However, that test takers are more familiar with a topic does not necessarily mean that they will perform better on them. In this regard, the literature is mixed. Test takers in the Powers and Fowles (1998) study did no better on topics they preferred. When the British Council's English Language Testing System (ELTS) was being revised towards becoming the IELTS, the plan was to divide test takers into six discipline areas and asked to write on topics specific to their field. This plan was abandoned when it was found that there was no systematic difference in test-takers' performance when responding to general and field-specific prompts (Hamp-Lyons, 1990). On the other hand, Tedick (1990) reports that ESL graduate students at three different ability levels all did better on a topic specific to their field than on a general topic. The prompts used in the study might be worth looking into, however. The general prompt is as follows, followed by the field-specific prompt:

> In a recent news magazine, a famous educator argued that progress makes us lazy. Do you agree or disagree with this point of view? Explain why you believe that progress does or does not cause people to become more lazy or passive. Support your answer with specific reasons and examples.
>
> Every field of study has controversial issues. Debate over these issues often occurs among professionals in the field and leads them to conduct research in order to look for evidence to support one position on the issue over another or others. Choose a current controversial issue in *your* [italics in original] field of study. Discuss the controversy and explain your position on the issue, being sure to provide examples to support your opinion. (p. 127)

It can be seen that the study was set up to maximally emphasize difference. The general prompt is on a subject people can probably write about even if they have not necessarily thought about it; in that way, it appears to fairly represent prompts such as are found in standardized writing assessments. The "specific" prompt, ironically, is the more general prompt. The field-specific prompt is virtually unconstrained, leaving respondents plenty of leeway on what to write about. That the topic is controversial issues means that there are already two or more fairly well-sketched out positions on the matter. It is not difficult to imagine that people will have more to say about the latter than the former. Add the fact that the subjects in this study are graduate students, who are steeped in their particular fields, and significant findings are clearly not a surprise.

In their study on the subject matter of prompts, Lee and Anderson (2007) split the difference. Similar to Tedick's (1990) study, their study involves placement of international students into ESL courses. In their case, the test is integrated; each writing prompt consisted of a 10-minute videotaped lecture combined with a 2-page reading article. Topics in the study included "trade barriers," "brain hemispheres," and "ethics." That is, unlike in Tedick, the field-specific prompts were topic limited rather than unconstrained. Controlling for language ability, as measured by TOEFL scores, Lee and Anderson found that while different topics did affect performance, students' majors were not related to writing performance. That is, background knowledge was not the factor that caused students to score differently on the different prompts. As the actual prompts are not provided, it is difficult to speculate on what the differences between the prompts might be.

Where subject matter is concerned, the purpose and context of assessment clearly matters. If the purpose is to assess facility in communicating on subjects where test takers have expertise, and not on writing ability in other areas, then a prompt such as in Tedick's (1990) would probably be appropriate. On the other hand, if the purpose is to assess writing ability more generally, then prompts that "bias for best" might not necessarily be desired. If test takers come with differing levels of ability, a "specific" topic such as in Tedick might well be biased against those segments of the population who do not have a particular area of expertise. From the discussion on the specific topic turning out to be the more general topic, one way of classifying topics, perhaps more properly belonging under rhetorical task specification, emerges: those that allow one to respond in one specific way (e.g., Do you agree or disagree regarding x?), and those that allow multiple possibilities (e.g., Give an example of y.). These can perhaps be called constrained and unconstrained prompts.

Weigle (2002) notes the surprisingly small amount of research on subject matter as a factor, and speculates that this might have to do with the infinite universe of possible subject matters, among other things. She concludes that

> while it seems sensible to assume that test takers can perform better when they are writing about subjects they know and care about than when they are not, it is likely that the effects of content are mitigated by other task variables. (p. 67)

Rhetorical Task, Grammatical Person, and Patterns of Exposition

In the literature, little work has been done focusing on the effects of different patterns of exposition on writing performance outcomes. Studies on the type of writing called for in a prompt have by and large compared personal versus impersonal writing, two of narrative, expository, and argumentative writing, or both comparisons in combination with each other. While results of the studies go in different directions, the overall picture that emerges is that, contrary to expectation, test takers tend to get lower scores on personal, narrative writing than on impersonal, argumentative writing, and that this is possibly caused by an interaction between task, test taker, and rater.

As mentioned, few studies have focused on the patterns of exposition that particular prompts elicit. Some students in Polio and Glew (1996) were advised by their teachers to select comparison and contrast prompts because these were supposedly easier; there is, however, no evidence to support this advice in the assessment literature. A few studies have compared traditional essay tasks with data commentary tasks (i.e. describing information found in graphs)—and perhaps these two task types might invite different patterns of exposition (Carlson, Bridgeman, Camp, & Waanders, 1985; Park, 1988; Reid, 1990; Weigle, 1999). Park found, for example, that Chinese and English language background test takers who majored in "hard" sciences did better on a task describing information in a graph compared to a traditional essay task, which was not observed among those with social science majors. There is, in other words, an interaction. The findings of these studies in general are that these two task types resulted in different linguistic production (Reid, 1990), that inexperienced raters were more severe in rating data commentary tasks, perhaps due to unfamiliarity with the format (Weigle, 1999), but that correlations between the two types of tasks were generally high (Carlson, et al., 1985). That the two task-types studied are rated differently only by inexperienced raters seems to imply that, with rater training, scores on the two types of tasks can be comparable, thereby providing some evidence that the two kinds of tasks tap the same underlying ability.

It is worth asking if the different kinds of production elicited by the two types of tasks are so different that traditional essays alone do not capture the whole construct of writing ability; if so, there clearly are implications for the proper assessment of writing. And there is a whole range of different patterns of exposition—process, cause and effect, classify, define—that have not yet received attention, likely because of their underrepresentation in testing practice. The writing test from which the current study's data come from is among those that do not cover the range of exposition patterns, and so this particular question will have to be left unaddressed.

Turning now to the larger body of research comparing different rhetorical specifications for traditional composition-type exams, a number of studies investigated performance on prompts that invited a personal, first person response versus those that called for impersonal, third person responses (Brossell & Ash, 1984; Greenberg, 1981; Hamp-Lyons & Mathias, 1994; Hinkel, 2002; Hoetker & Brossell, 1989; Spaan, 1993; Yu, 2007). Of these studies, Brossell and Ash, Greenberg, and Hoetker and Brossell found no significant differences. It appears that the lack of a finding can be attributed, in these cases, to the cue being so subtle that test takers were not likely to pick up on them. Here, for example are the sample prompts for personal and impersonal from Greenberg (1981, p. 94-95):

> In most American colleges, students must pass required courses in English, math, and science before they are allowed to take courses in their major areas of study. Instead of making all students attend all of their required courses, colleges should offer more independent study programs in which students could complete some of their courses on their own, working at their own pace. Do you agree or disagree with this statement? In an essay of about 300 words, explain and illustrate your answer in detail.
>
> In most American colleges, students must pass required courses in English, math, and science before they are allowed to take courses in their major area of study. Instead of making all of you attend all your required courses, colleges should offer you more independent study programs in which you could complete some of these courses on your own, working at your own pace. Do you agree or disagree with this statement? In an essay of about 300 words, explain and illustrate your answer in detail.

In the case of Hoetker and Brossell (1989), though, while there was no difference in the scores of compositions written in response to personal and impersonal prompts, the prompt did influence whether test takers wrote in the first or third person, and a separate ANOVA showed that raters gave significantly higher scores to first person essays than third person essays. One study, that of Spaan (1993), found that test takers performed better on narrative/personal prompts, though she offers that this might have been brought about by one of the argumentative/impersonal prompts being inaccessible to test takers: "What is your opinion of mercenary soldiers (those who are hired to fight for a country other than their own)? Discuss." (p. 101). It should also be noted that performing "better" in this case meant a difference on average so small that individual test-takers' final scores would have been the same. The other studies comparing personal versus impersonal (Hamp-Lyons & Mathias, 1994; Hinkel, 2002; Yu, 2007) found that the former received lower scores.

These findings need to be compared with those that investigate the difference among narrative, expository, and argumentative compositions. With the exception of the possible confound in Spaan (1993), the studies come down on the side of lower scores for narrative writing and higher scores for argumentative writing (Hamp-Lyons & Mathias, 1994; Quellmalz, Capell, & Chou, 1982; Wiseman, 2009). Quellmalz, et al., in a well-controlled multi-trait, multi-method study of eleventh and twelfth grade writers, found that students received significantly lower scores on narrative prompts than on expository prompts. Wiseman looked at a college writing placement test and had the same findings. Similarly, contrary to their expectations, Hamp-Lyons and Mathias found that argumentative/public compositions were scored higher than expository (narrative/descriptive)/private compositions in their sample of MELAB test takers.

There are possible explanations for these mixed findings. Quite a few (e.g., Hamp-Lyons & Mathias, 1994; O'Loughlin & Wigglesworth, 2007; Spaan, 1993) have speculated that raters are internally adjusting their rating behavior depending on how difficult they perceive a task to be. Thus, higher standards are applied to narrative compositions because they are perceived to be easier. An alternate explanation offered by Wiseman (2009) is that raters expect lower level test takers to choose the "easier" narrative prompts, and thus give lower ratings, befitting their expectations of test-taker abilities. Still another explanation is offered by Hake (1986), who found that straight narratives are more often mis-judged than expository compositions. The reason, Hinkel (2002) speculates, is that the easier and the more familiar a topic, as in personal/narrative writing, the responses also tend to be simpler. In other words, writers do not always demonstrate the full range of their abilities in such kinds of writing. Conversely, when forced to write on impersonal topics, Hinkel found that ESL essays tend to exhibit more native-like language features. Similar findings are seen in Crowhurst (1980), who found longer t-units in argumentative writing than in narrative writing among first language writers, in Yu (2007), who found that impersonal topics resulted in higher lexical diversity, and in Spaan (1993), who found that second language writers use more multi-syllabic words in argumentative/impersonal compositions.

Thus, one way of making sense of the data with regard to rhetorical task, at least as suggested by the literature, is to understand the interactions that can happen between the task, the writer, and the rater. A change in the rhetorical task prompted can elicit varying displays of test-taker ability, which possibly is further moderated by raters reacting to these different types of texts differently (Weigle, 1999). Still, it should be remembered that while different rhetorical tasks might elicit different kinds of linguistic production (Crowhurst, 1980; Ginther & Grant, 1997; Hinkel, 2002; Hoetker, 1982; Reid, 1990; Spaan, 1993), it does not necessarily follow that the rating received by the same person on different tasks will be different; that is a contention for research to investigate (Carlson, 1988; Fulcher & Reiter, 2003).

Prompt Wording and Specification

The wording of prompts, and whether changes to them affected outcomes, has received a decent amount of attention. Hinkel (2002) argues that prompt wording matters because students insert language from the prompt into their essays. For their part, Moore and Morton (1999) advise that prompts avoid asking test takers "should" questions, as it supposedly encourages hortatory writing, which goes against the balanced kind of writing required in higher education. Similarly, it has been pointed out that the word "discuss" has multiple meanings and can call for different kinds of writing (Evans, 1988; Horowitz, 1991; Weir, 2005). On a related note, a study has found that phrasing the task as a question or as a statement does not make a difference in outcomes (Brossell & Ash, 1984).

A number of studies have looked into the amount of information provided in the prompt. Kroll and Reid (1994) divide prompts into three categories: bare prompts, framed prompts, and text-based or reading-based prompts. The first is stated in relatively direct and simple terms (e.g., Do you favor or oppose x? Why?); the second presents a situation or circumstance, and the task is in reference to this; and the third has test takers read texts of some length and then asked to interpret, react to, or apply the information in those readings. For his part, Brossell (1983) divides roughly the first two categories into prompts that have low, moderate, and high information load. Brossell found that a medium level of specification resulted in longer essays and higher scores, though differences were not significant overall. In O'Loughlin and Wigglesworth (2007), tasks with less information elicited more complex language, but did not affect scores.

Test takers do consider the generality and specificity of prompts in their decision-making when allowed to choose (Polio & Glew, 1996; Powers & Fowles, 1998), and have also been shown to prefer shorter prompts (Chiste & O'Shea, 1988). This has not been to their advantage, though:

> Shorter, simple declarative sentences may appeal in their brevity but ultimately offer less insight into an essay's development and structure. Longer topic sentences… provide more direction even as they frighten away the less able student. (Gee, 1985, p. 84, qtd. in Chiste & O'Shea, 1988)

The consensus appears to be that a medium level of specification is ideal. Underspecified prompts require time and effort to narrow down, whereas very long prompts cause test takers to rely heavily on language and ideas in the prompt. A medium level of specification helps test takers focus without overloading them with information (Brossell, 1983; 1986; Lewkowicz, 1997). As the MELAB includes both bare and framed prompts, ranging in length from one to five sentences, the question of prompt specification is certainly worth investigating.

Another approach to classifying prompt specification is by counting the number of tasks the test taker is asked to complete. Kroll and Reid (1994) provide this example prompt which, by their reckoning, asks the test taker to do 13 different things:

> Some students believe that schools should only offer academic courses. Other students think that schools should offer classes in cultural enrichment and opportunities for sports activities as well as academic courses. Compare and contrast the advantages and disadvantages of attending a school that provides every type of class for students. Which of these types of school do you prefer? Give reasons and examples to support your choice. (p. 238)

The 13 tasks in the prompt are identified as follows: identify the advantages and disadvantages of (1, 2) each choice (3, 4); compare and contrast (5, 6) the advantages and disadvantages of (7, 8) each choice (9, 10); choose one of the choices (11) and give reasons and examples for the choice (12, 13). They provide a similar prompt that calls on test takers to do just two things:

> Some students want to attend schools which concentrate on academic courses only. Other students choose schools that require courses in music and art and participation in sports activities as well as in academic fields. Which of these types of schools would you prefer to attend? Use specific reasons to support your choice. (p. 238)

The claim here is that the larger the number of tasks required, the more difficult a prompt would be. However, this might not in fact be the case, as there is some evidence that both examinees and raters do not pay very much attention to whether all tasks in a given prompt are fulfilled, thereby rendering it a non-factor (Connor & Carrell, 1993).

Prompt Choice

In the preceding sections, prompt-inherent characteristics that could affect test performance were identified. In this section, the review considers an aspect of the writing test setup that could also conceivably affect outcomes. The MELAB presents each test taker with two prompts, and asks them to respond to just one. To an extent, this question about the provision of choice falls outside the scope of this dissertation, which limits itself to those aspects where test takers receive systematically different treatment; all test takers are given this choice. However, as this question is intricately connected to topics of interest in this study, raising important practical, theoretical, and measurement issues, it is given consideration here.

It is unclear if choice is a positive thing or not. On the one hand, the time available for completing the writing task is limited, and choice can create anxiety and take away time that would otherwise have been spent writing (Gabrielson, Gordon, & Englehard, 1995). In this regard, the research shows that test takers overestimated the amount of time they think they spent choosing a prompt, and that in reality they chose a prompt relatively quickly and proceeded to writing (Polio & Glew, 1996). On the other hand, most think that test takers like having a choice; they can respond to the prompt they think they can do better on.

The Effects of Choice

Whether test takers indeed do better when allowed to choose is debatable. A number of studies—in fields other than language assessment—indicate that by and large they do perform better under the choice condition, though the size of the effect varies, and though some also chose incorrectly and favored prompts that disadvantaged them. One approach asked test takers in Advanced Placement (AP) United States history and European history exams to indicate a preferred prompt but respond to both (Allen, Holland, & Thayer, 2005; Bridgeman, Morgan, & Wang, 1997). Overall, scores were a third of a standard deviation higher on preferred prompts. In both tests, test-takers' scores on the preferred prompt also correlated more highly with scores on an external criterion measure. On the other hand, approximately 30% of students in each exam performed better on the prompt they did not prefer. Differential performance has also been observed in English literature (Bell, 1997) and AP Chemistry tests (Fremer, Jackson, & McPeek, 1968; Wainer, Wang, & Thissen, 1994), in one case showing a five and a half point difference in test-taker performance between two prompts on a ten-point scale. It also appears that there are significant interactions between choice and test-taker gender, ethnicity, and level of proficiency. Male test takers appear better able to choose easier items, and less-able test takers apparently compound their problems by choosing what are actually the more difficult prompts for them (Wang, Wainer, & Thissen, 1995; Wainer & Thissen, 1994).

In studies of writing assessment specifically, effects were less clear. Chiste and O'Shea (1988) did find some pairing and order effects—test takers preferred those prompts that were shorter and that were listed earlier—though there was no correlation between choice and success. Powers and Fowles (1998) compared test-takers' judgments on prompts and their performance on the same on the GRE; they found only a weak and inconsistent relationship. Correlations between test-takers' ratings of prompts and their essay scores were significant in only two out of six cases. Overall, the more test takers preferred a prompt, the more their scores increased, though this finding was not statistically significant, possibly owing to sample size issues. Jennings, Fox, Graves, and Shohamy (1999), for their part, randomly assigned test takers on the Canadian Academic English Language assessment to a no-choice condition and a choice among five prompts condition. Holding ability level constant, similar to Powers and Fowles, test takers in the choice group had higher—though not statistically significant—scores than those under the no-choice condition. Textual analysis of the compositions also showed no difference among the two groups.

Effects in the opposite direction have also been found. In a study of eleventh-graders' persuasive essay writing, where half of the sample was assigned to each condition, the MANOVA showed that choice only had a small effect on the quality of essays, with those in the no-choice condition doing slightly better, but that there was a significant effect by gender and race (Gabrielson, et al., 1995). Another study, of basic writing, found that the more that test takers liked a prompt, the lower the scores they got (Powers, Fowles, Farnum, & Gerritz, 1992).

Jennings, et al. (1999) raise the possibility that the effects of choice might be attenuated. While the pool of prompts test takers can respond to is large, the element of "choice" is usually limited to choosing one of two prompts, e.g., Gabrielson, et al. (1995), Hamp-Lyons and Mathias (1994), Powers, et al. (1992), Spaan (1993). In the Powers, et al. study, for example, test takers indicated their preferences for 20 different prompts, but could only choose among two when it came time to write. Thus, it could well be that they were not able to write on their most preferred prompt, and differential performance on different prompts become more difficult to detect.

There might indeed be attenuation of the choice effect. On the other hand, when one compares the findings of studies on choice in the subject areas and of studies on choice in writing assessment, the differences remain striking. In both groups of studies, the amount of choice is limited. But then, studies of choice in the subject areas showed mostly significant positive topic effects, whereas studies of choice in writing assessment turned out mostly non-significant or mixed. A way to make sense of the disparate findings would be by thinking about the role of topic in each kind of test. It would make sense that specialized topic knowledge is more important in the subject areas—indeed, is part of the construct (Bachman & Palmer, 1996)—and test-takers' opinions about their doing better on one topic over another represent an informed judgment of topics they know and do not know. By contrast, as has been argued earlier, topic knowledge does not form part of the construct for generalized language proficiency exams. Test-takers' judgments about how well they would do on particular topics might not be as accurate or matter as much in writing assessments, since topic knowledge is not what they are being rated on. This difference in the role of topic in each kind of test has to be the best explanation for the different findings in studies of choice effects.

Theoretical and Measurement Considerations

Perhaps one reason more studies are not available is that in many cases, the way choice is set up creates measurement difficulties (Bradlow & Thomas, 1998; Wainer, Wang, & Thissen, 1994; Wang, Wainer, & Thissen, 1995). If test takers are asked to choose one out of two prompts to respond to, for example, non-equivalent test forms are created, and there is the possibility of selection bias. That is, two groups of test takers who differ in some characteristic or fashion might each prefer a different prompt. In this case, it becomes impossible to determine if any observed differences are due to the prompt or to whatever quality differentiates those two groups; there is a confound. Such is actually the case with the MELAB, and steps taken to deal with this issue in this study are detailed in the next chapter.

In any event, the discussion in the previous section shows that there are two possibilities: that choice makes a difference in test outcomes, and that choice makes no difference in outcomes. If choice does make a difference in test outcomes, it would seem to indicate that the prompts are tapping somewhat different abilities or somewhat different aspects of the same underlying ability. In which case, it would mean that test takers are either being tested on different things, or they are being measured on a limited part of the construct of interest. That is to say, there would be a problem with either construct validity or with construct under-representation (Messick, 1989; 1993). On the other hand, if choice does not make a difference in test outcomes, then it begs the question why choice is being offered, especially considering the measurement issues it creates noted above. Thus, Wainer, Wang, and Thissen (1994) conclude that "choice is only fair when it is unnecessary" (p. 197).

At least, that is, from a psychometric point of view. In Jennings, et al. (1999), whose study found no significant choice-related effects, test takers still indicated that choice is important to them. That is to say, even if choice made no difference in outcomes, it could make a difference in terms of alleviating test taker concerns, perhaps resulting in a more positive emotional affect during the test, and more of what has been called face validity. As Spaan (1993) writes, "choice may prove beneficial from an affective standpoint but remain neutral from the standpoint of performance or scoring" (p. 115).

Prompts and Test-taker Characteristics

The review has considered characteristics according to which prompts can vary, as well as prompt choice as a possible factor affecting outcomes in writing performance assessment. Investigations of test-taker characteristics that could interact with those prompt-related factors have focused on their gender, language background, and proficiency level; those are now given attention here.

Gender

Most of the studies that have looked into the relationship between writing prompt and test-taker gender have been conducted by the Educational Testing Service on their suite of exams. In a review summarizing the findings of studies for five different test programs, Breland, Bridgeman, and Fowles (1999) found differential item functioning (DIF) ranging from 0.07 to 0.14 in standard deviation units, all in favor of female test takers. Interestingly, on multiple-choice exams, the opposite was found and differences of roughly the same magnitude favored male test takers. The authors caution that the direction and size of the differences are highly sensitive to sample selection, and the findings should not be generalized beyond those exams studied. In a newer study on TOEFL computer-based test writing prompts, Breland, Lee, Najarian, and Muraki (2004) also found that women did better than men on some prompts. Only three of 87 prompts in this regard had differences greater than 0.2 of a standard deviation, with the largest being 0.24. Where gender differences in difficulty obtained, a panel of experts looked at examinee responses and hypothesized that women did better on topics dealing with music and the arts, housing and living conditions, and human relationships. Broer, Lee, Rizavi, and Powers (2005), for their part, looked into the GRE using several DIF detection techniques. Their findings also show that prompts favored women by a small amount, with the difference being largest at the higher ends of the scale for Issue prompts and at lower ends of the scale for Argument prompts.

In their study of persuasive writing among 11th graders, Gabrielson, et al. (1995), found significant effects for gender and race both of which were larger than the effects of task and choice. Differences favored women, with effect sizes ranging from 0.35 to 0.45 on the four traits the essays were rated on, and with slightly larger effect sizes under the choice condition rather than the no-choice condition. Results were mixed in Willingham and Cole (1997), who found some prompts that favored white women and some prompts that were easier for white men in an Advanced Placement English language and composition test. That is to say, there was non-uniform DIF. Finally, Park (2006) followed the logistic regression procedures (Zumbo, 1999) used in many of the ETS studies to investigate DIF in 10 MELAB writing prompts. The findings were that the few prompts that were initially flagged for uniform and non-uniform DIF had very small effect sizes and could not be considered as showing DIF. However, because of sample size issues, these findings cannot be said to be definitive. On the whole, it would appear that where gender is a factor, the differences tend to be small and tend to favor female test takers.

Language Background

Studies have considered the different production of writers from different language backgrounds on different tasks (Park, 1988; Reid, 1990). Reid, for example, studied the performance of writers whose first languages were Arabic, Chinese, English, and Spanish on a comparison and contrast task and a graph/data commentary task. She found that there was a greater percentage of content words for all groups on the comparison and contrast task and, conversely, also greater pronoun usage for all groups on the graph task. There were also non-uniform effects, however. Writers from three of the language backgrounds, with the exception of the Spanish group, showed greater fluency on the graph task. There was also greater use of passive-voice verbs in the comparison and contrast task for Arabic and Chinese writers, but not for English and Spanish writers. In Park's study, differences in production depending on language background and area of academic specialization. The question now would be if these differences in production also resulted in differences in outcomes in ways not related to the construct.

A number of the studies that looked into the relationship between prompt and gender have also investigated the relationship between prompt and language background. As with the findings on gender, Park (2006) reported the absence of significant differences. Broer, et al., who studied the GRE, compared test takers for whom English was their best language against test takers whose best language was not English. They found a moderate-sized difference in favor of the former. And the review of Breland, et al. (1999) considered the performance of ESL Hispanics and Asian Americans on three different exams. Prompts favored White Americans by 0.72 and 0.76 standard deviation units, respectively, which is substantially larger than the differences observed with gender. Their more important finding, however, might be the observation that differences related to race were smallest on essay tests than on tests using other methods and formats (e.g., multiple choice).

Other studies that have looked into language background as a factor include Lee, Breland, and Muraki (2004), who compared test takers with Indo-European and East Asian first languages. That is, where the comparison groups in other studies have been English first language people, this study compares two groups of non-native English writers. In this study, two-thirds of all prompts did not show any DIF, while those that did showed generally small uniform and non-uniform DIF. Where non-uniform DIF was seen, the differences became smaller, and sometimes switched, at the highest ability levels. On the whole, the differences between the two groups were largely attributable to differences in English language ability, which is to say that the prompts show not item bias but item impact (Clauser & Mazor, 1998; Penfield & Lam, 2000; Zumbo, 1999); differential probabilities of success are likely because test takers actually differ in the ability of interest.

Taking language background as a factor, there is a notable difference in findings depending on what the comparison group is; that is, whether the population includes those for whom English is a first language. DIF is more likely to show up when the English first language test takers are included. Thus, the purpose and the population of test takers for a given exam are important considerations, at least where this particular factor and interaction is concerned. On another note, it is worth repeating that DIF was found to be smallest in performance assessments than in exams using other test methods, providing some support to the notion that the former is a fairer way of measuring test-takers' writing abilities.

Proficiency Level

A test-taker's language ability might also partially determine whether prompts are or are not a factor in writing assessment. Studies that have considered this interaction are unanimous in showing that prompts are more of a factor among test takers at lower proficiency levels. In Spaan's (1993) study, subjects were divided into beginning, intermediate, and advanced levels according to their reading and listening scores on the MELAB. While tests for significance were not conducted, it can be seen that beginners' scores on the narrative/personal prompts and argumentative/impersonal prompts differed by 1.71 points, narrowed to 0.78 among intermediate-level test takers, and was further reduced to 0.03 for the advanced group. (It might also be worth noting that the former two groups received higher scores on the narrative/personal prompts, whereas the opposite was true for advanced learners.) Lee, et al. (2004), who compared test takers from Indo-European and East Asian language backgrounds, found that where non-uniform DIF existed, that language group membership had effects at low levels of language proficiency but not at higher levels. They attribute this finding to the lower-level test takers being more likely to resort to their first languages, which of course differ from English to different degrees.

There is also apparently an interaction between prompt, proficiency level, and choice. Test takers are known to consider the perceived ease or difficulty of prompts in choosing what to write on (Chiste & O'Shea, 1988; Mohan & Lo, 1985; Polio & Glew, 1996). However, it appears test takers at different ability levels are not equally good at making the choice, with lower level candidates more likely to incorrectly choose the more difficult prompt (Gabrielson, et al., 1995; Jennings, et al., 1999; Pollitt &

Hutchinson, 1987; Wainer & Thissen, 1994). Jennings, et al., noting the different choice patterns of low and high-proficiency test takers, speculate that the former group might have had more difficulty in reading the prompts, thereby impairing their ability to choose the best prompt for them. If this is so, then, at the minimum, special attention needs to be given to the wording of prompts—which was discussed earlier—to make sure that these prompts are as accessible as they can be, and not hamper lower-level test takers from providing a fair demonstration of their abilities.

Raters as a Factor

Raters are instrumental in performance assessments, as they are the link between test-takers' performances and the scores that these performances are awarded. While certain test providers have begun introducing automated rating by computers (e.g., the use of e-Rater by the Educational Testing Service), they also recognize that scoring by computers alone is not feasible—writing is, after all, a communicative activity between human beings, and how another person receives and perceives a piece of text remains beyond the capability of computers—and the scoring process of major English language writing proficiency tests always still involves human raters.

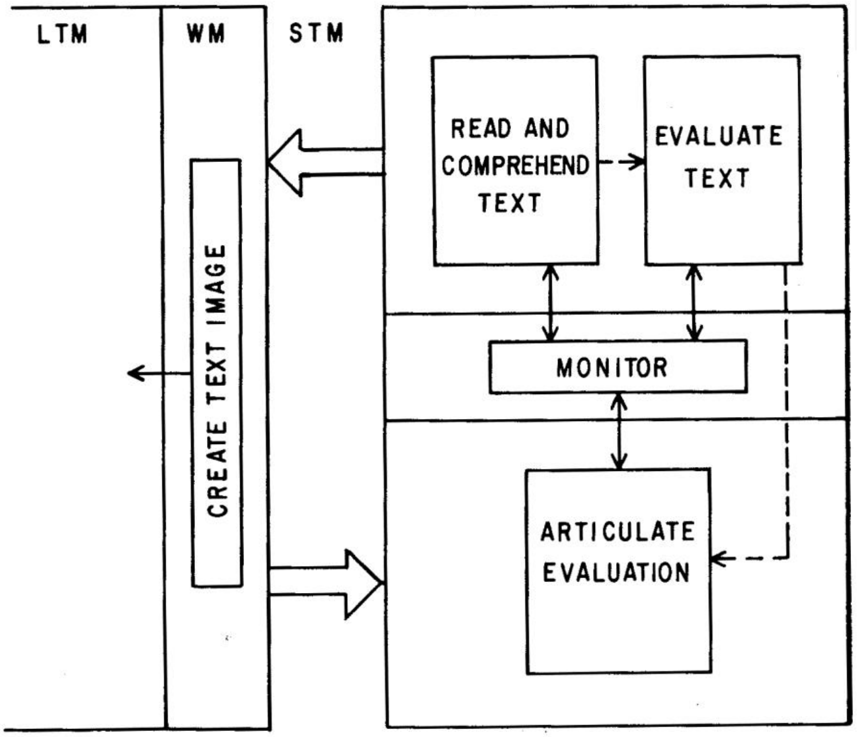
However, the use of human raters also brings with it potential issues such as subjectivity and reliability, which could in turn affect the validity of test scores. An early study, for example, found that of 300 essays scored on a nine-point scale, 94% received at least seven different scores, and an inter-rater reliability of 0.31 (Diederich, French, & Carlton, 1961). Through better specified rating scales and other measures, it has since become possible to achieve much higher levels of inter-rater reliability. However, the desirability of increasing agreement by and of itself has come under question (Connor-Linton, 1995a; Lumley & McNamara, 1995; Reed & Cohen, 2001; Weigle, 1998). The inter-rater reliability statistic only says something about the product of assessment but not about the process, and "if we don't know what raters are doing… then we don't know what their ratings mean" (Connor-Linton, 1995a, p. 763). Raters could well be agreeing on things that have nothing to do with what is being measured. Thus, there is the need to better understand the rating process itself—how raters go about the task of rating and what factors they actually consider—as well as the rater characteristics that could affect raters' rating behavior.

It is to studies of these that the chapter now turns. First, it will review a body of mostly qualitative studies, the majority of which utilize think-aloud protocols (Ericsson & Simon, 1993), that investigate raters' decision making processes. From these studies are developed models of the rating activity. Following, studies of rater characteristics that are thought to influence rating behavior, mostly using quantitative methodology, are looked into. Interactions between raters and prompts are also considered.

The Process of Rating

Raters can have different approaches to rating, or what is called "reading styles" (Milanovic, Saville, & Shen, 1996). While this might indicate difference, there is general agreement that the same underlying basic process is being followed in rating performance assessments. Freedman and Calfee (1983), coming from an information processing perspective, present a model where raters (1) read and comprehend text, (2) evaluate the text, and (3) articulate their evaluation (Figure 3.1). (In their model, it has to be noted, raters create a text image after reading and comprehending, and it is that text image, rather than the text itself, that raters evaluate and store impressions of.) While acknowledging the possibility that rating could be a linear process, Freedman and Calfee believe that it is more likely one that is recursive, where chunks of text are evaluated as they are read and comprehended. The monitor in their model allows raters to revise their evaluations as they read and evaluate more pieces of text.

Figure 3.1 Freedman and Calfee's (1983, p.92) Model of the Rating Process



Similarly, Cumming, Kantor, and Powers (2002) propose that the prototypical sequence of decision making involves three steps: (1) scanning the composition for surface level identification, (2) engaging in interpretation strategies, reading the essay while exerting certain judgment strategies, and (3) articulating a scoring decision, while summarizing and reinterpreting judgments. Finally, Lumley (2006) offers that the basic process includes (1) reading and prescoring, (2) scoring, and (3) revising and finalizing the rating.

The three studies above differ in their particulars. For example, Lumley's involved experienced raters who were working with a rating scale, while raters in Cumming, et al. were both experienced and inexperienced, and were also not provided with rubrics. Freedman and Calfee's model was the result of a think-aloud study, as the other two were, but of experimental studies. But different as the particulars of the studies were, they come to remarkably similar conclusions about what the rating process looks like. The only notable difference is Cumming, et al.'s introduction of a pre-reading stage, where raters look at such features as format, length, and paragraphing before actually reading the compositions. Otherwise, the three studies agree that raters read, understand, and evaluate compositions, and then articulate their evaluations. They all have elements in their definitions acknowledging that evaluation is a recursive activity. In Freedman and Calfee, the arrows that form a feedback loop; in Cumming, et al., the hint that evaluation begins in the second step and that "summarizing and reinterpreting" happens in the third; and in Lumley, the mention of prescoring, scoring, and revising.

Raters' Rating Behaviors

Building on work by Cumming (1990), Cumming, et al. (2002) elaborate on the basic process by identifying different rating behaviors. In their study, 17 raters rated TOEFL writing tasks, and their think-alouds were coded according to the rating behaviors their statements reflected. These behaviors were then organized into a descriptive framework that can be said to include the universe of rating behaviors, at least for this group of raters for this type of writing task (Table 3.3). The framework divides behaviors into two. Interpretation strategies refer to the ways raters try to understand compositions, while judgment strategies are those used by raters to evaluate and rate said compositions. These strategies are also divided according to what raters focus on, whether monitoring their own behavior, considering ideas in the compositions, or paying attention to language in the compositions.

Table 3.3 Descriptive Framework of Decision-Making Behaviors while Rating TOEFL Writing Tasks (Cumming, et al., 2002, p.88)

Self-Monitoring Focus	Rhetorical and Ideational Focus	Language Focus
Interpretation Strategies		
Read or interpret prompt or task input or both	Discern rhetorical structure	Classify errors into types
Read or reread composition	Summarize ideas or propositions	Interpret or edit ambiguous or unclear phrases
Envision personal situation of the writer	Scan whole composition or observe layout	
Judgment Strategies		
Decide on macrostrategy for reading and rating; compare with other compositions; or summarize, distinguish, or tally judgments collectively	Assess reasoning, logic, or topic development	Assess quantity of total written production
Consider own personal response or biases	Assess task completion or relevance	Assess comprehensibility and fluency
Define or revise own criteria	Assess coherence and identify redundancies	Consider frequency and gravity of errors
Articulate general impression	Assess interest, originality, or creativity	Consider lexis
Articulate or revise scoring decision	Assess text organization, style, register, discourse functions, or genre	Consider syntax or morphology
	Consider use and understanding of source material	Consider spelling or punctuation
	Rate ideas or rhetoric	Rate language overall

The identification of rating behaviors allowed Cumming, et al. to assess differences among raters, e.g., raters who have different language backgrounds and rating experience, providing insights into rater selection and training. On the other hand, there are caveats and questions that remain. The first has to do with think-aloud protocols as a methodology. Raters themselves have offered that think-alouds not only provided a partial view of their thinking, as some aspects of the rating activity are deeply intuitive and difficult to verbalize, they also in some ways altered raters' rating processes (Barkaoui, 2007b; Lumley, 2006). Second, as was previously mentioned, the study by Cumming, et al. was not done in an operational context, and raters were also not provided a rating scale. How would the provision of a rating scale change rater behavior? Which behaviors in the framework will become irrelevant? And which other behaviors will be observed? Finally, the study does not tell us how behaviors lead to raters giving particular ratings. For example, let us say that a rater employs those judgment strategies that focus on language features. How did the rater decide that a composition deserves a mark of three instead of four? Other studies attempt to answer this question.

Raters' Decision Making

A study by Erdosy (2003) homes in on the decision-making process. From among the raters in Cumming, et al. (2002), four were chosen and asked to construct scoring criteria while assessing a sample of TOEFL essays and were later interviewed. The finding was that all four raters bridged performance and proficiency by internalizing a developmental trajectory for language learners, and then they determined test-takers' proficiency by locating test-takers' positions in that trajectory. Thus, there is some evidence for scoring validity, in that scores are in fact based on some construct of language ability.

On the other hand, the four raters in this study had differing definitions of proficiency and also constructed different developmental trajectories. The data indicate that raters' language backgrounds and teaching experience were factors that resulted in these differences, making these factors worth exploring. But in any case, what this indicates is that the raters were each judging a different construct, and Erdosy concludes with the assertion that

> although training in assessment procedures can enable a group of raters to render reliable judgments using a particular rating scale, only raters who, largely because of similarities in their teaching experiences, have shared attitudes toward the acquisition of language proficiency… are likely to base their judgments on a shared construct of writing proficiency. (p. 57)

That raters are unlikely to judge based on a shared construct might be too hasty a conclusion to make for a study that in fact asked raters to construct their own rating scales. What the study in fact does not address is whether raters from different backgrounds—whether language, academic, or professional—can be trained to share a common understanding of language proficiency and language development not of their own creation, and be trained to rate according to it appropriately. As Alderson (1991) argues, "at some point it is necessary to decide: 'these are the scales, and this is how they are to be interpreted. If you cannot agree with that, you cannot be certified [as a rater for a particular test]'" (p. 82). If this common understanding is possible is what needs to be established.

A study that focuses on raters' decision making process where a rating scale is provided is that of Lumley (2002, 2006). As with Erdosy, four raters were studied, but in this case, all four were trained and experienced in rating the Special Test of English Proficiency (STEP), a test used by the Australian government for immigration-related decisions. Raters rated according to the multiple-trait STEP rating scale. Further, half the ratings were done under operational conditions, and the other half while doing the think-aloud, so that the effects of using think-alouds can also be assessed.

To the question whether raters can be trained to a common understanding of a scale, Lumley finds that "to a considerable extent the raters do interpret the categories and levels in similar ways" (p. 237). The exact meaning of this, however, needs to be glossed. Lumley notes that raters do refer to the text vis-à-vis the scale,

> but we do not get clear statements about exactly *how* this relationship is made. The assessment operates automatically, at a sub-conscious level, using a mass of simultaneously processed information, and the thought processes remain largely inaccessible to us as well as to the raters themselves. (p. 237)

Clearly, there are limits to the think-aloud procedure, but from what the raters did say, it is possible to see that the rating scale does not actually come first for these raters, nor does it sit at the center of the scoring process. Rather, the rater reads and develops a sense of the text; the features they notice and observe are massive in number but not necessarily easy to organize or articulate. This is where the rating scale comes in:

> The value of the scale is to guide that process, by channeling the complexity of the raters' sense of the text into something simpler, more manageable and, ultimately, more reliable, which the testing institution can use as evidence for its claim about test takers' ability. (p. 197)

And also:

> [T]he role of the scale is as a classificatory scheme for the raters' impression of the texts. Out of the mass of features they notice in any text,

> the scale guides what they are to say about it, thus assisting raters in *articulating* these impressions. (p. 237)

That is to say, it is not so much that raters come to a common understanding of what the rating scale means, but that the scales provide them with a common vocabulary for describing and talking about what they have read and their impressions of it. The rating scale, which in and of itself is inert, is brought to life by the rater's engaging with it. From this, it can also be seen that rating is not a simple task of recognition or analysis of features. Rather, it involves "squeezing, shaping, defining, arbitrating, comparing, and rejecting" (p. 240), and is a complex, problem-solving activity.

DeRemer (1998) also argues that rating is a problem-solving activity. But where Lumley's raters appear to approach the rating task in the same way, conscientiously considering both text and scale, DeRemer finds that her raters have different "task elaborations" or approaches to the rating activity. These are general impression scoring, text-based evaluation, and rubric-based evaluation. Several other studies (Sakyi, 2000; Vaughan, 1991) also identify different rater approaches to rating, but these can all be classified under DeRemer's categories (Table 3.4). Sakyi and Vaughan also both note that when in doubt, raters fell back to one or two particular features to help them arrive at a final rating.

Table 3.4 Rater Approaches to Rating

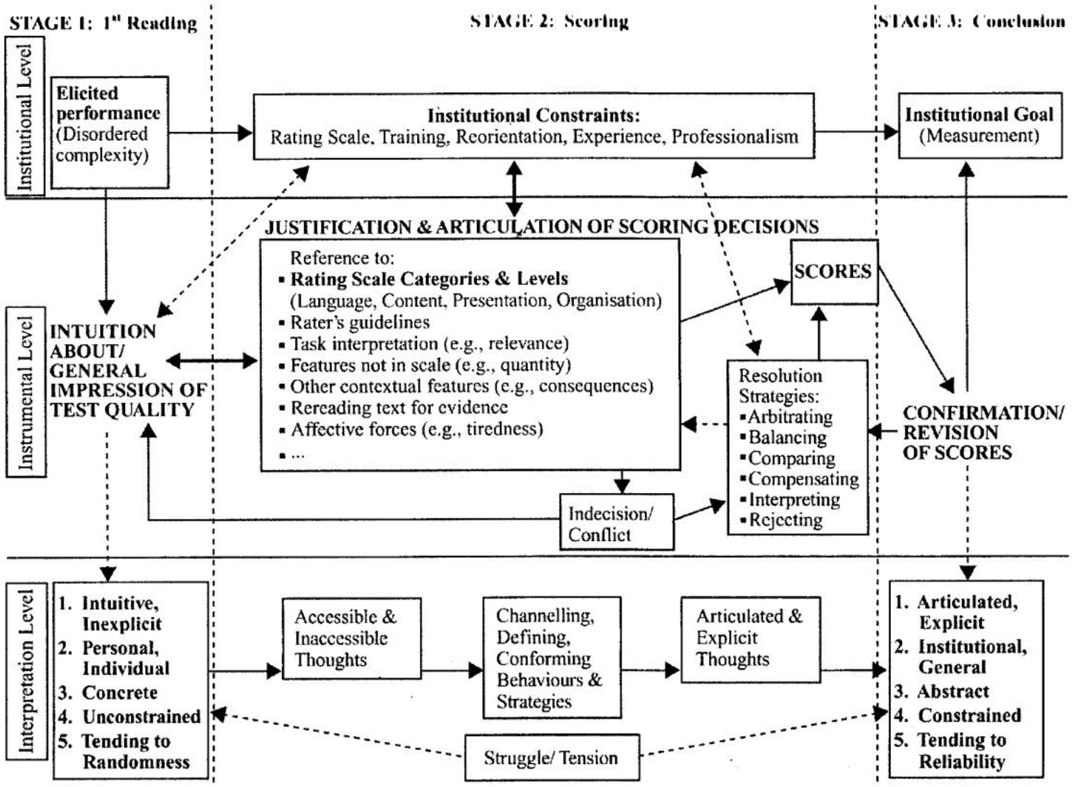
DeRemer (1998)	♦general impression	♦ text based	♦ rubric based
Sakyi (2000)	♦ personal reaction	♦ errors ♦ topic and idea presentation	♦ scoring guide
Vaughan (1991)	♦ first impression dominates ♦ “laughing rater”	♦ single focus ♦ two category ♦ grammar oriented	

From these studies, it would appear that raters consider just one aspect while rating, to the exclusion of others. That, however, does not sound very plausible. It is doubtful, for example, that a rater could go by general impression alone and give an accurate score without having considered the text or the rating scale at all. The text must have in some way contributed to the impression, unless the rater was going purely by surface features such as handwriting and such. Similarly, the rater must have seen the rating scale at some point in the past; it could well be that having worked with the scale for a long time, the rater makes no explicit mention of the scale even as he or she incorporates information from it, compares it with the general impression, and arrives at a rating for a text. That is to say, more likely, the raters’ thinking-aloud in these three studies emphasized one aspect or the other, even though their actual rating decisions were actually partly based on all three aspects. The studies’ observation that in complicated cases raters’ resorted to one or two features to help them make a decision would support this argument, in that the going back to some criterion or criteria means that other criteria had in fact been considered but deemed inadequate for making a decision. Lumley’s (2006) work, which engaged with raters more extensively, probably presents a more complete picture of the rating process. The general impressions and the textual features noticed and identified by raters are myriad, and the rubric provided them with language for articulating their observations.

Model of the Rating Process

At the conclusion of his study, Lumley (2006) offers what is arguably the most complete model of the rating process to date (Figure 3.2). As has previously been discussed, there are three basic stages to the rating process: reading and prescoring, scoring, and revising and finalizing of scores. It can also be seen that this process takes place on three different levels, what Lumley calls the institutional level, the instrumental level, and the interpretation level.

Figure 3.2 Lumley's (2006, p. 291) Model of the Rating Process


STAGE 1: 1st Reading
STAGE 2: Scoring
STAGE 3: Conclusion
Institutional Level
Elicited performance (Disordered complexity)
Institutional Constraints: Rating Scale, Training, Reorientation, Experience, Professionalism
Institutional Goal (Measurement)
JUSTIFICATION & ARTICULATION OF SCORING DECISIONS
Instrumental Level
INTUITION ABOUT/ GENERAL IMPRESSION OF TEST QUALITY
Reference to:
▪ Rating Scale Categories & Levels (Language, Content, Presentation, Organisation)
▪ Rater's guidelines
▪ Task interpretation (e.g., relevance)
▪ Features not in scale (e.g., quantity)
▪ Other contextual features (e.g., consequences)
▪ Rereading text for evidence
▪ Affective forces (e.g., tiredness)
▪ ...
SCORES
Resolution Strategies:
▪Arbitrating
▪Balancing
▪Comparing
▪Compensating
▪Interpreting
▪Rejecting
CONFIRMATION/ REVISION OF SCORES
Indecision/ Conflict
Interpretation Level
1. Intuitive, Inexplicit
2. Personal, Individual
3. Concrete
4. Unconstrained
5. Tending to Randomness
Accessible & Inaccessible Thoughts
Channelling, Defining, Conforming Behaviours & Strategies
Articulated & Explicit Thoughts
1. Articulated, Explicit
2. Institutional, General
3. Abstract
4. Constrained
5. Tending to Reliability
Struggle/ Tension

The instrumental level, showing rater behaviors, is the most visible part of the process and, perhaps for that reason, what most studies have focused on. At the beginning of the process, raters read compositions and develop an intuitive impression of their quality. (The “text image” in Freeman and Calfee’s (1983) model would be a rough analogue to this.) This stage is important because even though a rating has not been awarded, a judgment about texts has been made. For this reason, this stage is also called the prescoring stage. As to the source of those impressions, studies suggest that raters’ backgrounds play a part (Cumming, 1990; Cumming, et al., 2002; Pula & Huot, 1993). These background factors will be discussed in the next section and are among the factors that the present study seeks to explicate.

In the scoring stage, raters refer primarily to the rating scale, constantly going back and forth between the stated criteria and the intuitive impressions they have formed of the texts, as indicated by the thick double-headed arrow. Other than the rating scale, however, raters also consider additional guidelines and features not in the scale. It can be seen that raters’ interpretation of the task is one such consideration, which could bring out the prompt effect that this dissertation investigates. Following reading, rereading, and providing justification, raters may proceed to giving a rating or a score. On the other hand, they may experience conflict and indecision, in which event they resort to a range of resolution strategies such as arbitrating (between different components of the scale) or comparing (a composition with some other), among others suggested by the literature (Cumming, et al., 2002; Weigle, 1994). At this point, raters may refer back to the rating scale and to other features in the big box, as well as their impressions of the text. This process will at some point lead them to settle on a rating for the composition.

Raters then move on to the final stage of the process, where they confirm the ratings they have given, or go through a process of revising a rating until they are satisfied, in which case the process is completed. From the model, it can be seen that the rating scale does not have primacy, as is often thought, but rather, that rater and rating scale are both at the center of the process; it is the rater who develops a sense of the text and then channels those impressions towards a rating, aided by the rating scale but also by a host of other factors and strategies. Contrary to DeRemer (1998), who argued that raters have different main foci, whether their general impression, the text, or the scale, this model shows that all three are in some way considered and reconciled, with different emphasis depending on occasion, en route to a rating. And contrary to Charney (1984), who argued that ratings can be reliable only when the reading is done quickly and superficially, this model shows that the process is one characterized by complexity and thoughtful consideration of texts (cf. Huot, 1993).

Underlying the instrumental level is the interpretation level. This is the level where tension and struggle exists. Raters are first of all people who have their natural ways of reading, which are personal, intuitive, and unconstrained. But as they are reading in a particular context, as raters for a particular purpose and provided a rating scale, they need to select, arrange, and channel their diverging thoughts, so that they converge into something explicit that can be articulated and which conforms with the requirements. Thus, what is inexplicit and personal, concrete and tending towards randomness, becomes something explicit and institutional, abstract and tending towards reliability.

Finally, while most studies have focused on the cognitive aspects of the rating process, it can be seen that there is an institutional component to the activity. There is always some institutional reason why writing performance is elicited. At this first stage, performances can be said to exhibit disordered complexity, as they are not in a form where inferences can be drawn and decisions made based on them. For this reason, a series of institutional constraints are introduced so that performances can be described with consistency. These constraints include first of all the rating scale, but also include the selection of raters with the right kinds of experience and professionalism, as well as the provision of rater training and retraining as necessary. The operationalization of these constraints leads to the goal of the institution, which is measurement. The filters put in place turn disordered complexity into something standardized and reliable, a score, which the institution can use as a basis for decisions it needs to make.

The Behavior and Characteristics of Raters

As Lumley's (2006) model of the rating process shows, raters are an important component to the rating process. The rating process involves tension and struggle, as raters are people who come to the task of rating with different personalities and histories. And as these rater characteristics and backgrounds inform their rating, it is important to know what effects these have on the ratings they give. In the following sections, the chapter reviews the literature with regard to raters' general rating patterns and tendencies, the effect of training and increased rating experience, and whether cultural and language background affect their rating performance. One other aspect of raters' background that has been considered in the literature is their profession. This literature has considered the difference between people who have and have not been trained in language and education (Huot, 1993; Schoonen, Vergeer, & Eiting, 1997; Shohamy, et al., 1992), and of language teachers and teachers in different content areas (Brown, 1991; Cumming, et al., 2002; Janopoulous, 1992; O'Hagan, 1999; O'Loughlin, 1992; Santos, 1988; Song & Caruso, 1996; Weigle, Bolt, & Valsecchi, 2003). However, as all MELAB raters share the same professional background—that of assessment professionals—the current study is not able to inform that aspect of scholarly investigation. Those aspects that it can are now considered.

General Tendencies and Consistency Over Time

The literature on rating quality indicates there are four major categories of rater errors (Engelhard, 1994; Saal, Downey, & Lahey, 1980). The first is the tendency towards severity or leniency. That is, a rater consistently gives lower or higher ratings than a performance deserves. Engelhard (1994) offers that it is perhaps best to see raters as being on a continuum of severity and leniency. If test takers are rated by raters who vary much in severity, then either some people are getting higher scores than they deserve or some are getting lower scores than is appropriate, which could clearly affect the validity of these scores. In addition, other than overall severity, raters could also differ in severity at different points in the scale (Hill, 1996; Schaefer, 2008) where, for example, it might be relatively easier to get a rating of four with one rater, but harder to get a rating of six. A solution to such problems is available, provided the raters involved are consistent in their severity. When that is the case, within the Rasch (1980) approach to measurement, it is possible to measure how much more severe or lenient they are than appropriate, and then to make adjustments to the scores of test takers they rated accordingly. In addition to individual raters differing in severity, several studies have looked into the relative severity and leniency of groups of raters of writing, such as by rating experience (e.g., Weigle, 1998; 1999) and by language background (e.g., Hill, 1996; Kondo-Brown, 2002). These will be discussed in detail in later sections.

The second category of rater error is the halo effect, where raters do not distinguish between different aspects of a composition when there are indeed differences among them. This error applies only to performance assessments that employ multiple-trait scoring, and as such is not within the purview of the present study. Central tendency is the third kind of rater error, where a rater mostly uses middle of the scale and is reluctant to use the two ends of the scale. Among other things, this creates an artificial sense of consistency to the raters' ratings overall. The fourth rater-error category is closely related to the third. Restriction of range refers to the extent to which ratings are able to discriminate different test takers into different performance levels. If test-takers' performances are not differentiated, then the purposes of measurement are defeated. Clearly, raters showing central tendency are more likely to also result in restriction of range. As with severity, there are measures within Rasch methodology that can identify these problems. Fit statistics can show overfit, which would be an indication of central tendency. The distribution of test-takers' abilities across the ability range, and analysis of the rating scale, can show whether there is a restriction of range.

The four categories of errors identified above are cross-sectional, i.e. they consider rater performance at one particular time. Many assessments, however, are not just given once, but many times across periods of time. Studies have thus also looked into rating consistency over time. Fitzpatrick, Ercikan, Yen, and Ferrara (1998) conducted two studies where exams of third, fifth, and eighth grade students across a range of subject areas were rescored after one year. They found that the absolute standardized mean differences were generally small, in the range of one-tenth to two-tenths of a standard deviation. One of the exceptions, though, involved writing in Grade 5, where the mean difference can be considered large. The authors also calculated the correlations between total scores in the first and second sets of ratings. Correlations were consistently highest in mathematics, and consistently lowest in writing. Pearson correlations for third, fifth, and eighth grade writing were 0.58, 0.59, and 0.72, respectively. In this study, however, the raters in time one and time two were not the same people.

Cho (1999) had ten raters read the same 20 student essays four times, with an interval of four to six weeks between readings. The study found high Kendall's tau-b correlation coefficients, with most raters registering internal consistency values higher than 0.7 across comparisons. Cho wonders, however, whether there might have been a "memory effect," which presents a possible confound.

For their part, Congdon and McQueen (2000) examined the ratings of 16 raters in seven rating sessions over a period of nine days, where performances rated on the first day were re-rated by the same raters on the last day. In this case, raters read an average of 173 essays each day, thus making it less likely that they would remember what scores they had given to essays they read more than once. On a day to day basis, it was found that ratings for the group became more stable beginning with the fourth rating session. It suggested to the authors that a period of practice and getting used to the task was necessary for these raters who only had a half-day training session. The rating sessions were also divided by a weekend when no ratings took place. The finding of a “weekend effect” suggested that re-training was necessary for these raters. As for the first and last day, when the same compositions were rated and re-rated, it was found that nine raters were significantly more severe and one rater significantly more lenient. The difference in average rater severity between the first and last day for all raters was 0.45 logits, or approximately 0.14 of a score point. A difference of one score point, in the context of this study, was the equivalent of one year’s progress in elementary school. How meaningful a difference 0.14 of a score point is is for the relevant authorities to determine. That is to say, differences will always be found, and the question is how large can differences become before they are deemed unacceptable. In the end, it still requires and comes down to a matter of human judgment.

Rater Training and Experience

Studies comparing novice and experienced raters indicate that there are differences between them. One difference is in the way they go about rating. Huot (1993), in a think-aloud study, found that while novice and experienced holistic raters considered the same criteria, their reading process was quite different. Novice, in this case, referred to raters who were not given scoring guidelines. To begin, novice raters in his study tended to make more comments as they read, whereas expert raters made more comments after they had finished reading. Compared to novice raters, expert raters also made a greater percentage of personal comments. The reason for these differences, the study shows, is that expert raters already knew what to evaluate in a composition and had already developed a strategy for rating. Their strategies were not all the same, but each had a strategy that worked for them. (Cumming, et al. (2002) also have the same finding in this regard.) Knowing the criteria and possessing a strategy allowed expert raters to not have to focus so much on particulars as they read, and allowed them to engage with the texts on a more personal level, and then to evaluate the compositions more generally and as a whole after they had finished reading. By comparison, novice raters were apparently attempting to develop judging criteria at the same time that they were reading, the reason they made more comments as they went along. Their attempt to do several things at once resulted in more remarks having to do with the steps they were taking (139 for novices, 4 for experts), and it is not surprising that most novice raters reported that their rating technique broke down at some point. It must be said that the novice raters arrived at the same criteria as the expert raters, but the attention devoted to discovering the criteria meant they were not able to engage with the texts and consider them more holistically. Huot's (1993) findings have been replicated in a more recent study by Barkaoui (2009).

Wolfe (1997; Wolfe, Kao, & Ranney, 1998), using a somewhat different approach, confirms many of Huot's (1993) findings. Unlike in Huot, all the raters in these studies were provided a rating scale. Depending on how highly their ratings correlated with others, raters were classified as competent, intermediate, and proficient. Like Huot, Wolfe and his colleagues found that proficient raters, compared to the others, had fewer interruptions while reading and were able to withhold judgment until after they had finished reading. They also made more general comments, rather than on specific pieces of the text. As well, proficient raters considered all features equally and used more rubric-related language.

What these studies suggest, then, is that training might not have the same effect on all raters. Or alternately, that the strategies of successful raters identified here need to form a part of rater training; rater training should help raters develop a sense of the standards, as well as develop ways of approaching the rating task.

Having considered the way raters of different experience and expertise rate, it is appropriate to consider whether their rating performance indeed differs. In a study by Shohamy, Gordon, and Kraemer (1992), ten trained and untrained raters each read 50 compositions and rated them on three scales: holistic, communicative, and accuracy. Further, each group of ten raters included five people with a background in language and education and five people without. While inter-rater reliability coefficients were generally high overall, ranging from 0.80 to 0.93, not surprisingly, the trained raters achieved higher reliability coefficients than the untrained raters (0.91-0.93 vs. 0.80-0.90). Having a background in language and education made no difference in rating quality. Weigle (1998), for her part, had eight experienced raters and eight inexperienced raters rate writing samples on a college placement test, in a pre- and post-training design. Using multi-faceted Rasch methodology, she found that inexperienced raters were more severe and less consistent in their rating compared to experienced raters. Training reduced but did not eliminate the differences in severity between the two groups of raters. The consistency of inexperienced raters, however, showed much improvement after training. The conclusion of the study is that rater training helps improve intra-rater rather than inter-rater reliability.

There are suggestions in the literature that there is an interaction between training/experience and writing task. Another study by Weigle (1999), also using a pre- and post-training design, had experienced and inexperienced raters rate compositions that responded to two different tasks: one task was more like a traditional essay, calling on the writer to make and defend a choice; the other task asked the writer to interpret a graph. Results showed that before training, inexperienced raters were more severe in rating the graph task. However, this difference in severity disappeared after training. Accompanying think-aloud protocols indicated that the two tasks elicited compositions that were differently structured, and that the scoring rubrics were not as easy to use in scoring the graph task, which accounts for the inexperienced raters' difficulties with the graph task at first in Weigle's study. The current study only had test-takers' responses to one task type, so it cannot investigate this interaction. It can investigate, however, whether experience—operationalized by length of tenure—is a predictor of rating performance for this group of trained raters.

Language and Cultural Background

Using the native speaker as the point of reference has had a long and contested history in applied linguistics (Davies, 2004). It is no surprise that studies have compared native and non-native raters and considered the appropriateness of the latter serving as raters. To some, non-native speakers serving as raters are considered an exceptional category, if not downright unacceptable. On the other hand, there are those who argue that depending on the context of the testing situation, non-native raters are actually more appropriate for evaluating test taker performance (Hill, 1996).

Different reasons have been forwarded for why native and non-native speakers might differ in their rating behavior. One is the notion of contrastive rhetoric (Grabe & Kaplan, 1996; Kaplan, 1966; Leki, 1991), that people from different cultural backgrounds have different cultural preferences, making them prefer different rhetorical patterns. People can also come from cultures with very different norms surrounding communicative events, e.g., politeness (Brown, 1995). In one study, non-native raters noticed and flagged biased prompts, which the native-English speaking raters did not (Erdosy, 2003). Studies in cultural psychology (e.g., Nisbett, 2003) have similarly found differences in how people from different cultures think and what they are likely to notice and pay attention to.

A study by Zhang (1998) found that native-English and native-Chinese speakers differed in their expectations of writing related to rhetorical patterns, including overall organization, use of supporting evidence, use of conjunctions, register, objectivity, and persuasion. The Chinese teachers focused on accuracy, whereas the American focused on intra- and inter-sentential features. Kobayashi and Rinnert (1996) investigated how readers from different backgrounds evaluate compositions containing different rhetorical patterns. While no difference in overall assessment was observed, they found that

> Japanese students who had not received English writing instruction preferred the Japanese rhetorical pattern, native English teachers favored the American rhetorical pattern, and Japanese students who had received English writing instruction and Japanese teachers valued features of both patterns. (p. 397-8)

In a follow-up study (Rinnert & Kobayashi, 2001), analysis showed that the more ESL instruction Japanese students received, the closer their perceptions were to their native-English speaking teachers. In the opposite direction, native-English speakers have also been shown to alter their reading of non-native English texts with experience (Hamp-Lyons, 1989). Thus, it would appear that differences in preferred rhetorical patterns are not absolute and can be altered with exposure.

There can also be more directly linguistic reasons why native and non-native raters might differ. Focusing on the evaluation of English in particular, there are many who now recognize International and World Englishes (Hamp-Lyons & Davies, 2008; Kachru, 1992), and raters can come from parts of the world with well-developed varieties of English, which can differ from a standard dialect in significant ways (Lowenberg, 2000). A non-native rater can conceivably read and consider as acceptable certain features that would not be acceptable in standard dialects of English, and thus rate more leniently. In the context of international assessments of English proficiency, which tend to measure some standard dialect or dialects of English, that might cause them to rate in ways considered inappropriate by the examination provider.

Many of the studies that have considered the severity and leniency of raters from different language backgrounds have focused on speaking (Barnwell, 1989; Brown, 1995; Fayer & Krasinski, 1987; M. K. Kim, 2001; Y. H. Kim, 2009; Lumley & McNamara, 1995; Gass, Winke, & Reed, 2007). Their findings have been mixed. In any event, it appears that quite a few of the findings might not apply to writing. In Fayer and Krasinski's (1987) study, what raters found irritating were errors in pronunciation and hesitation—features of language performance absent in writing. Similarly, Brown's (1995) raters treated test takers differentially on politeness, which is more salient in speaking than in writing.

A study that has considered rater severity in writing is that of Shi (2001), where raters considered Chinese learners of English. The study found that teachers who were Chinese and non-native speakers of English identified more negative features of learners' writing while native-English speaking teachers made significantly more positive comments. However, when the teachers scored the learners, an inversion was observed: the native-speaker raters who noted more positive features gave lower marks than did the non-native speaker raters who dwelled on the negative aspects of the writing. Shi attributes this phenomenon to the raters' taking on a "double role of a strict native speaker and a lenient EFL teacher" (p. 312). Another study, of Japanese EFL and American ESL teachers, showed that while the two groups of teachers' ratings were comparable, they focused on different aspects of writing: the Japanese teachers on accuracy, the American teachers on intra- and inter-sentential features (Connor-Linton, 1995b).

The context of the above studies is clearly different from that of international English assessments, where raters do not have the dual role noted by Shi (2001). In addition, most studies on rater language background effects thus far have compared native and non-native raters from only one other language. The possibility of a language distance effect on language examination performance has been mooted (Chiswick & Miller, 2004; Elder & Davies, 1998). It is theorized, for example, that Japanese is at a greater distance from English than is Spanish. It could thus well be that on an English language examination, the amount of bias for or against Japanese could be larger compared to Spanish. This question can only be answered if multiple languages are accounted for at the same time.

An attempt to account for both rater language background and language distance effect was made by Hamp-Lyons and Davies (2008). Their study looked at MELAB compositions written by native speakers of Arabic, Bahasa Indonesia/Malay, Chinese, Japanese, Tamil, and Yoruba. In addition to the official MELAB ratings, given by native English speakers, these compositions were also rated by raters who shared the examinees' first language and by raters who did not share their first language – this to see whether there is language background-related bias in the exam and among raters with differing first language backgrounds. Their study, however, had a number of intervening variables—trained and untrained raters with differing levels of reliability, the use of two different rating scales, and a data set of limited size, among other things—making it difficult to draw conclusions regarding language background-related bias. The issue of rater language background thus requires further investigation.

Raters and Prompts

The two main variables this study investigates are raters and prompts, and it should be asked whether a relationship exists between the two or not. Rater expectations apparently have an effect on ratings (Barritt, Stock, & Clark, 1986; Diederich, 1974; Freedman, 1984; Stock & Robinson, 1987), and several studies have given their authors cause for wondering whether raters are adjusting their rating behavior according to their sense of particular prompts. In Hamp-Lyons and Mathias (1994), reviewed earlier, the experts all agreed that prompts calling for argumentative/public writing are more difficult than those calling for expository/private writing. However, they found that responses to the former kind of prompts actually received higher scores than those that responded to the latter. Spaan (1993) had the same unexpected finding. Thus, Hamp-Lyons and Mathias (1994) write:

> [W]e must consider the possibility that essay readers are consciously or unconsciously compensating in their scoring for relative prompt difficulty based on their own, internalized, difficulty estimates…. [S]uch a compensatory mechanism would tend to negate the expected effect of prompt difficulty on scores. (p. 59-60)

Along the same vein, O'Loughlin and Wigglesworth (2007) compared outcomes on prompts containing more and less information, and found no difference in final scores. This finding could just mean that differences in prompt content made no difference in writing and rating outcomes, but the authors also could not help but speculate whether raters were compensating for a task that, containing more information, appeared to be more difficult. These speculations about an interaction between rater and prompts they perceive to be more difficult have never been formally examined.

Research Questions

This study seeks to discover how the validity, reliability, and fairness of a second language performance assessment are affected by the assignment of different prompts and different raters to different test takers. In light of the literature, a number of research questions are posed regarding prompts:

1. Consistent with the requirements of test validity, reliability, and fairness, to what extent are the writing prompts in a large-scale English language proficiency examination comparable in difficulty, and to what extent does the test reflect the absence of a prompt effect?

2. To what extent can it be shown that there is no prompt effect related to topic domain, rhetorical task, prompt length, task constraint, expected grammatical person of response, or number of tasks?
3. To what extent are writing prompts not differentially difficult for test takers of different genders, language backgrounds, and proficiency level?

Three questions about raters are also posed:

4. Consistent with the requirements of test validity, reliability, and fairness, to what extent do raters in a large-scale English language proficiency examination rate appropriately and consistently, and to what extent does the test reflect the absence of a rater effect?
5. To what extent can it be shown that there is no rater effect as a result of experience, time, and language background?
6. To what extent can it be shown that raters do not alter their rating behavior depending on perceived differences in prompt difficulty or perceived proficiency-related prompt selection behavior among test takers?

In each set of three questions, the first is a more general question, while the second and third questions consider more specific aspects of prompts and raters respectively. It follows that the answers to the general questions depend in part on the answer to the more specific questions.

Chapter Summary

This chapter identified those aspects of prompts and raters that might have an effect on the scores that performance writing samples receive. Prompts can differ according to topic domain, rhetorical task, length, degree of constraint, number of tasks, and expected grammatical person of response. There might also be interactions between prompts and test takers of different genders, language background, and proficiency level. In addition, the provision of choice creates measurement issues that need to be addressed. Raters, for their part, can differ in their rating tendencies, in their experience rating, and in their language backgrounds. Whether they alter their rating behavior depending on the prompt or test taker encountered has also been raised. Based on the literature, six research questions were formulated; three each regarding prompts and raters. In the next chapter, I provide the details of a study designed to investigate these questions regarding the effects of prompts and raters on assessment outcomes.

CHAPTER 4

METHODS

In this chapter, I describe the specifics of this study, which is designed to answer the question: **How are the validity, reliability, and fairness of a second language writing performance assessment affected by aspects of the examination that are systematically varied for different test takers?** This question was operationalized into six research questions at the end of the previous chapter, of which, the two more general questions are:

- Consistent with the requirements of test validity, reliability, and fairness, to what extent are the writing prompts in a large-scale English language proficiency examination comparable in difficulty, and to what extent does the test reflect the absence of a prompt effect?
- Consistent with the requirements of test validity, reliability, and fairness, to what extent do raters in a large-scale English language proficiency examination rate appropriately and consistently, and to what extent does the test reflect the absence of a rater effect?

This chapter has three main sections. The first section presents the details of the data used in this study. It provides information about the prompts used in the MELAB writing test, about the raters who rate the test, and about the test takers whose writing are being evaluated. Because test takers are given a choice of prompts and respond to only one prompt, a possible confound is created. An approach to dealing with this problem is discussed. The second section of this chapter centers on the methods and analyses used by the study. It will describe the Rasch approach and its multi-facet extension, and how the approach is appropriate for the study's purposes. Finally, the third section details the procedures followed in answering each question. The chapter concludes with a discussion of the study's limitations.

Data

The nature and construct of the MELAB writing test was already described in Chapter 2. Here, I describe the prompts, raters, and test takers involved in the writing test between October 2003 and February 2008. This period of approximately four and a half years covers the available MELAB database up to the time the request for data was made.

Prompts

Because the MELAB is a secure test and because a large number of the prompts are still in use, this study cannot reveal the prompts or the exact way in which they are worded. Below, however, are a few retired prompts, which should be enough to provide a general sense of what the test's writing prompts are like:

> What do you think is your country's greatest problem? Explain in detail and tell what you think can be done about it.
>
> An optimist is someone who sees the good side of things. A pessimist sees the bad side. Are you an optimist or a pessimist? Relate a personal experience that shows this.

> In many countries, people are forced to stop working and to retire at age 65. Yet many of the world's most creative people did their best work after that age. Write about the advantages and disadvantages of forced retirement.

As can be seen, the prompts can differ in length, in topic domain, and in how personal a response they invite, among other dimensions. They are also not questions where there is a "correct" or "incorrect" response. None of the prompts involve graphs, data commentary, or other extra required readings. The prompts also appear to be generally accessible to educated adult test takers.

The study's data included 66 prompts similar to the above. Six of the prompts, however, were retired soon after the beginning of the period covered by the study and were taken by only a small number of test takers; where there are hundreds of compositions for each of the other prompts, for these six the number ranged as low as five and nine compositions. Any results related to these prompts, which only a very small number of test takers and raters encountered, are not likely to be very meaningful. For this reason, these six prompts were excluded from the study. The remaining 60 prompts ranged in length between 12 and 82 words, with a mean of 38.47 (Table 4.1). In terms of sentences, they were as short as a single sentence and as long as five sentences.

Table 4.1. Length of MELAB Writing Prompts

	Mean	SD	Min	Max
Words	38.47	14.72	12	82
Sentences	3.17	0.98	1	5

One of the study's questions concerns the degree to which differences in prompt length, topic domain, rhetorical task, task constraint, expected grammatical person of response, and number of tasks affects test outcomes. The length of prompts—in number of sentences—can easily be counted. The other dimensions of prompts, however, cannot be arrived at by mere counting. For these dimensions, the prompts were individually coded by me and by a testing professional with expertise in writing assessment according to the categories in Table 4.2. The categories for topic domain are those used internally by the ELI, while the categories for the other dimensions came out of the literature review in the previous chapter.

Table 4.2 Prompt Coding Categories

Dimension	Categories
Topic Domain	Business Education Personal Social
Rhetorical Task	Argumentative Expository Narrative
Task Constraint	Constrained Unconstrained
Grammatical Person of Response	First Person Third Person
Number of Tasks	1 to n

After the initial coding, the two of us had a reconciliation meeting where we discussed those cases where our coding did not match. While we each changed our coding on some items, we agreed that our lack of agreement on some prompts was really the result of those prompts covering multiple possible codes, (e.g., it can invite either a first person or a third person response), and did not constitute real disagreements. We chose to leave those “disagreements” as they were, rather than force an agreement that might actually misrepresent the nature of those prompts. Our agreement rates before and after our meeting are given in Table 4.3 while the codes for each prompt are given in Appendix C.

Table 4.3. Prompt Coding Agreement Rates, Percentages

	Topic Domain	Rhetorical Task	Task Constraint	Grammatical Person	Number of Tasks
Initial	92	83	75	85	85
After meeting	95	95	87	95	95

Raters

A small group of 24 raters rated compositions in the period covered by the study. The raters all went through the same standardized, multi-stage training and certification program. This involved work done independently and in collaboration with a trainer on the nature of the test in general, the rating scale and the benchmarks, and calibration activities. Afterwards, raters began monitored live rating, where they keep track of their agreement rates with other raters and receive feedback about their ratings. They became certified after reading a minimum number of live compositions to an acceptable level of rating quality. The raters were also all regular employees of the ELI of the University of Michigan at the time they served as raters. That is to say, all the ratings were done at a single site where people saw and worked with each other on a daily basis, and not by freelance contractors working in different locations. This arrangement increased the likelihood that raters had a shared understanding of the construct and of the rating scale, contributing to the validity of the scores (Erdosy, 2003; Purves, 1992). Indeed, high reliability and agreement rates are reported by the MELAB program among these raters, with low discrepant ratings ranging from 1.52% to 3.11% for the complete years covered by this study (Johnson, 2005, 2006, 2007; Johnson & Song, 2008). On the other hand, it means that caution needs to be exercised in generalizing the results of this study to test programs that employ freelance raters with different day-jobs working in different locations.

The rater-related variables that the study is interested in include experience and language background. Raters' experience rating MELAB writing (Table 4.4)—measured as the time when they started rating up to their last recorded rating in the data—ranged from two months to over 23 years, with an average of just over five years. The number of compositions read can also be a measure of experience, and is also given in Table 4.4, as is ratings per month of experience, which incorporates both measures of experience.

Table 4.4 Raters' Rating Experience and Rating Information

Rater	Ratings in This Data	Experience	Ratings in Data Per Month of Experience	Mean	SD
R01	6622	4 y 9 mos.	116.2	76.53	7.73
R02	13	2 mos.	6.5	76.54	6.59
R03	139	15 y 3 mos.	0.8	77.14	7.65
R04	98	2 mos.	49.0	75.00	10.95
R05	45	6 y 0 mos.	0.6	76.91	6.56
R06	762	2 y 2 mos.	29.3	75.70	7.11
R07	942	1 y 3 mos.	62.8	76.53	6.85
R08	158	2 y 6 mos.	5.3	79.09	7.54
R09	5153	3 y 6 mos.	122.7	77.15	6.65
R10	731	1 y 4 mos.	45.7	76.24	6.61
R11	2401	4 y 5 mos.	45.3	77.32	6.89
R12	1589	2 y 8 mos.	49.7	75.33	6.90
R13	145	10 mos.	14.5	74.59	8.63
R14	688	6 mos.	114.7	76.22	7.41
R15	232	3 y 0 mos.	6.4	77.09	6.89
R16	123	20 y 8 mos.	0.5	75.20	6.52
R17	113	2 mos.	56.5	74.63	7.97
R18	49	23 y 2 mos.	0.2	76.39	8.32
R19	133	2 y 5 mos.	4.6	77.83	6.92
R20	3368	2 y 5 mos.	116.1	76.48	7.47
R21	1712	2 y 8 mos.	53.5	75.98	6.67
R22	3240	2 y 8 mos.	101.3	77.01	6.66
R23	1236	3 y 6 mos.	29.4	77.76	7.43
R24	139	15 y 3 mos.	0.8	75.99	7.88

While most of the raters were native speakers of American English, four raters had first language backgrounds other than English. Raters R05 and R08 were originally from South America and have Spanish for their first language. R09 came from the Philippines, and whose home languages growing up were Chinese (Amoy) and Filipino (Tagalog). Finally, R21 is a native speaker of Korean born in Korea, but who grew up mostly in the United States. All four are highly proficient in English.

Another rater-related question of this study has to do with the possibility that raters internally adjust their ratings based on perceived difficulty of prompts (Hamp-Lyons & Mathias, 1994; O'Loughlin & Wigglesworth, 2007; Spaan, 1993) or on perceived proficiency level of test-takers, i.e., that lower-level candidates are more likely to choose narrative prompts, and so these prompts are awarded lower scores (Wiseman, 2009). To the end of exploring these possibilities, I developed two forms for raters to fill out, soliciting their perceptions of particular prompts. The instructions given to the raters are given in Appendix D. No additional information was given to raters regarding the purpose of the forms or the questions they were intended to answer.

Form A (Appendix E) asked raters to respond to the question "Compared to the average prompt in the pool of MELAB writing prompts, is this prompt easier, about average, or more difficult to get a high score on?" and to indicate their perception of each prompt by checking the appropriate box on the 5-point Likert-type scale. Because asking raters to provide their perceptions of all 60 prompts might represent too much of a cognitive load and result in fatigue-related error, I decided to reduce the number by half, to 30. The 30 prompts were selected according to the following criteria: (1) all prompts for which there was agreement between both coders across all dimensions after the initial coding; (2) all remaining prompts inviting a narrative response; and (3) all remaining prompts inviting an unconstrained response for which there was eventual agreement across all dimensions. The first criterion was instituted so that prompts for which the coding are most reliable are included, while the latter two criteria were introduced so that codes of interest would be represented by an adequate number of prompts in the form. The order of the prompts was then randomized using a random sequence generator before they were put on the form.

For its part, Form B presented pairs of prompts, according to their original pairing in MELAB test forms, and asked raters to indicate for each pair if lower-level candidates are more likely to choose one than the other, and if so, which one. Whereas Form A asked raters to keep all prompts in mind while considering a particular prompt, Form B simply asked raters to compare two prompts. I believe this presented raters with much less of a cognitive load, and so included more prompts in this form. All 30 prompts in Form A, plus their respective pairs (if those pairs were not already in Form A), were included in Form B, resulting in 48 prompts or 24 comparisons. Because this form involves pairs of prompts as they are paired in actual test forms, and as responses are relative to individual pairs of prompts rather than to the total pool of prompts, randomizing their arrangement was not done as it was neither necessary nor desirable.

While there were 24 raters in the data, some of those raters no longer work for the MELAB program. Thus, Forms A and B were filled in only by current raters and by former raters who left relatively recently and who could be contacted, minus recusals by those closely connected to this study. In total, there were 10 respondents to the rater perception forms.

Involvement of the Researcher

I need to disclose that I, the researcher, am one of the raters included in the study. Like all other raters, I completed the training and certification program. All of the ratings I gave were in the normal course of employment, most of them at a time when I had no idea that I would undertake the present research. In any event, there was no reason for me to rate in any different fashion, and any attempt to do so would have been flagged by the MELAB program anyway. Thus, there should be no need to exclude the ratings I gave from the main part of the present study. Where Forms A and B are concerned, because I was the one who devised the instrument, there was the strong likelihood that I would respond to it in a different fashion than other raters would. Because that part of the study does not include all raters anyway, I chose to exclude myself from responding to the forms.

Test Takers

All test takers who took the MELAB between October 2003 and February 2008, and all the ratings assigned to their compositions, minus those who had missing data and those prompts with very few data points as mentioned above, were part of the study's sample. The resulting sample included 29,831 ratings for 10,536 test takers. Missing and excluded data totaled 1,233 or just below 4% of the sample. The largest number of those excluded was for lack of information about what prompt they wrote on, from a period of time when this information was systematically not entered into the MELAB database. Because prompts are rotated and have a roughly equal chance of getting assigned at any given time, the missing prompt data is also likely to be somewhat random. Given that knowledge, given the large remaining sample, and given the method of analysis used in this study, the missing data should not be a threat to the integrity of the sample or of the results. On another note, it should be remembered that each MELAB composition is rated by at least two raters, and a third rater adjudicates if the ratings of the first two differ by more than one scale point (ELI, 2005). And test takers are allowed to request a rescore if they feel that the score they received is inaccurate. Thus, there are potentially up to six ratings for each composition. Further, there are also test takers who take the test more than once; hence, the apparently large disparity between the number of test takers and the number of ratings.

Those who took the MELAB in this time period were between 14 and 80 years old, and had an average age of just under 29 years old (SD = 11.1). Female test takers were more numerous than male test takers, accounting for 57.29% of all test takers. The test takers came from more than 115 different first-language backgrounds. However, languages represented by less than 10 test takers were recoded under "other" categories by region, leaving 59 first languages. Those language and language groups accounting for at least one percent of the sample are given in Table 4.5. (Language group refers to languages which have multiple dialects, e.g., Amoy, Cantonese, Hakka, and Mandarin were all coded under "Chinese"). It will be noted that there are a number of test takers whose first language is English, and for whom the test is not designed. These test takers potentially represent a different population than all the other test takers. Johnson and Lim (2009) have investigated this aspect of the MELAB data; their study showed that the only effect of including these test takers is an underestimation of English first-language test-takers' abilities. Estimates for all others are not significantly affected. Given those findings, the study chose to include English first-language test takers, with the caveat that findings related to those test takers be interpreted with appropriate caution.

Table 4.5 Well-Represented First-Language Backgrounds

Language	Number
Chinese	2248
Filipino	1259
Arabic	714
Farsi	670
Korean	542
English	438
Spanish	434
Punjabi	394
Russian	388
Urdu	372
Hindi	268
Romanian	222
Malayalam	173
Somali	164
Japanese	153
Gujarati	139
Bengali	120
Vietnamese	120
Portuguese	113
German	110

Sample and Sampling Issues

It was noted in Chapter 3 that the MELAB writing test asked test takers to choose between two prompts and to respond to just one, and that this provision created a possible confound. For each pair of prompts, there might be some characteristic of test takers—including ability level—that might make them more likely to choose one prompt over the other. Thus, one cannot know whether differences in outcomes are caused by differences in prompt or by that other, unidentified characteristic. In the language of the Rasch approach employed by this study, the 60 prompts would belong to 60 "disjoint subsets," and estimates for prompt difficulty may not be unambiguously compared. One solution for this problem would be to have a number of test takers respond to more than one prompt at a time. If these overlaps are appropriately designed, each prompt would be related to every other prompt through some common sample of responses. This idea was broached to MELAB administrators, but it was deemed that such a data collection effort was not feasible.

Matching Test Takers

Because that option was not available, the study opted for another solution suggested by the literature and made possible by the data. Other comparability studies have approached this problem by creating matching variables, primarily an overall English language ability variable based on test-takers' scores in skill areas other than writing (e.g., Breland, et al., 2004; Broer, et al., 2005; Lee, et al, 2004). Different test takers are then matched according to their similarity in this regard. This is arguably an imperfect solution, and depends on certain arguments regarding the similarity and difference between the different language skill areas. It can also mask differences between people whose total English language ability scores are identical, but who have different skill profiles.

Fortunately, compared to the above mentioned studies, the present study did not need to make assumptions that are quite that strong. Recall that in those studies, different people were being matched according to the variable and assumed to be comparable. The present data, for its part, happened to include a large number of test takers who—perhaps to obtain a higher score—took the MELAB more than once; in some cases, up to ten times (cf. Johnson, 2004). Especially for those who have taken the MELAB many times—that is, people who for whatever purpose they may have need a particular score but clearly have not attained it, as evidenced by their taking the exam over and over—an argument can be made that fossilization (Han, 2004; Selinker, 1972) has occurred, and that those candidates' abilities are approximately the same from one sitting of the test to another. Whether that was indeed the case was checked, to some extent, by looking at their scores on other sections of the MELAB (i.e. listening, reading) to see if they were comparable. Thus, unlike other studies where matching depended on similarities in test scores alone, the current study matched according to similarities in test scores *and* the fact that those being matched were in fact the same person, providing greater confidence that matches being made were warranted. When a match was made, the result was that the data now had one person who had responded to two writing prompts. With enough of those matches, the data became connected, permitting unambiguous comparisons between prompts.

For the study, I followed a procedure that minimized matching to the minimum amount necessary to achieve connection. I began by using the most restrictive conditions: a person's two sittings would be matched only if there was a small elapsed time between sittings for the test, and only if standardized part scores for listening and grammar, vocabulary and reading (GVR) were entirely identical. Because the MELAB requires six weeks between sittings for the test, which is almost two months, I began by setting the elapsed time criterion at three months. Running the data through the FACETS software under this condition, connection between prompts was not achieved. Neither was it achieved when I loosened the conditions to no more than three months between sittings, no more than one point difference between listening scores, and no more than one point difference between GVR scores.

Table 4.6 Connection of Prompts Under Various Matching Conditions

Match only if…			Connection achieved?
Elapsed time between tests ≤	Difference between listening scores	Difference between GVR scores	
3 months	0	0	no
3 months	≤ 1	≤ 1	no
3 months	≤ 2	≤ 2	yes
2 months	≤ 2	≤ 2	no
4 months	≤ 1	≤ 1	no

Connection among prompts was achieved when I set the parameters at no more than three months between sittings, no more than two points difference between standardized listening scores, and no more than two points difference between standardized GVR scores (Table 4.6). This involved a modest total of 214 test takers. For these 214 test takers, the absolute mean difference between their writing scores was 3.74, with a standard deviation of 3.11. Keeping in mind that the difference between each MELAB scale point is four or six (see Appendix B), this means that these test-takers' scores on different sittings for the test and different topics differed by less than one scale point. Having achieved connection with 3-2-2 as the parameters, I tried other combinations where I tightened the elapsed time criterion (2-2-2) or the score comparability criterion but allowing for slightly greater elapsed time (4-1-1). Prompts remained disconnected under those conditions. Thus, I adopted the combination of 3-2-2 as the basis for matching in the study, as it was the one that minimized matching (and assumptions) while creating connection. And those parameters are very much defensible, given that three months is a relatively small amount of time, and given that the allowed differences between part scores are less than the standard errors for each of those tests (SEM for Listening= 3; SEM for GVR=4). Apart from the above, all other instances where test takers took the MELAB multiple times were treated by the study as independent observations.

Using Extant Data

The issues discussed above reveal some of the weaknesses of what has been called convenience sampling (McMillan & Schumacher, 2001), or the use of data that just happens to be available. It has been argued that generalizations and conclusions from such data can be limited if not misleading, as particular sampling decisions are not known, which could systematically alter or bias outcomes. I would argue, however, that the data used by this study exhibits the characteristics of a population sample more than it does those of a convenience sample. Because the present data covers all prompts, all raters, and all test takers over a period of four and a half years, and after data preparation and cleaning kept over 96% of all data, one can have confidence that any findings would also apply to those who take the MELAB in other years, subject to the prompts and the raters being comparable. The present instance would only conceivably be a convenience sample if the population is defined as all similar, advanced-level, international, English proficiency tests. But certainly, the present study has no plans, nor should any other, of generalizing to that level, given that such tests, similar as they may be, have known differences in test methods, constructs, contents, and rater training methods.

By the same token, it is admitted that because this study does not use specially-collected data, except in a very limited way, there are certain issues and questions raised in the literature that the present study cannot answer. But then, there is no study that can answer every question. On the other hand, studies that collect data for the express purpose of answering certain questions frequently have particulars that differ in important ways from real-world conditions, thus leaving doubts whether they apply in real practice. The use of operational data, while possibly limiting the number of questions that can be investigated, provides more confidence about the applicability of its findings.

Method

Classical approaches to estimating prompt and rater effects have suffered from a number of weaknesses. The difficulty of a prompt, for example, was measured by the average score of people who responded to that prompt. That number, however, depended on the sample of people taking the test, and changed if a different set of test takers took the test. In the same manner, it was also impossible to compare two average scores and say how much more difficult one prompt was compared to another because the numbers were ordinal rather than interval measures, and because they were sample-dependent. One could not conclude whether differences were due to the prompts being different or due to the test takers being different. As well, it was a group-level measure, and one could not make meaningful statements about individual test takers based on it. Where raters are concerned, the conventional measure of rating quality has been inter-rater reliability. However, one could imagine two raters where one consistently gave a higher rating than the other, and by the same amount. In such an instance, the correlation between them would be perfect, but a test-taker’s score—and perhaps the test-taker’s passing or failing—would differ depending on the rater assigned. That is to say, inter-rater reliability only demonstrated the consistency of raters' rank ordering of candidates, but does not capture the relative severity and leniency of raters. It was also impossible to capture any bias or interaction effects.

To more accurately quantify prompt and rater effects, then, an approach is needed where (1) prompt difficulty is estimated separately from test-taker ability but still, following Bachman's (2002) arguments presented in Chapter 3, cognizant of and conceptualizing it as relative to test takers; (2) raters' leniency and severity are accounted for in addition to ability to rank order test takers; (3) measures of prompt and raters are interval measures, so that both order and distance can be known; (4) measures for prompts and raters are expressed on a common scale so as to make possible meaningful comparisons between them, and (5) the approach allows for the measurement of interactions among various aspects of these variables.

An approach that is known to provide these qualities and that has been used in studying performance assessments is the multi-facet extension of Rasch measurement (Linacre, 1989; Rasch, 1980). Studies using this approach have looked into the assessment of writing (e.g., Englehard, 1992; 1994; Hill, 1996; Kondo-Brown, 2002; Schaefer, 2008; Weigle, 1998, 1999) and the assessment of speaking (e.g., Brown, 1995, 2003; Lumley & O'Sullivan, 2005), as well as performance assessment in other areas such as the health professions (e.g., Fisher, 1993; Lunz & Stahl, 1990, 1993).

To be sure, other approaches have been proposed that possess many of the same qualities as multi-faceted Rasch, e.g., generalizability theory (Brennan, 2001; Shavelson & Webb, 1991), logistic regression (Swaminathan & Rogers, 1990; Zumbo, 1999). There are advantages to using multi-faceted Rasch, however. Generalizability theory, for example, while able to identify interactions, because of its focus on aggregation and generalizability, is not able to identify which individual raters and test takers are involved in said interactions (Bachman, Lynch, & Mason, 1995; Kozaki, 2004; McNamara & Roever, 2006). The identification of individual raters involved in interactions is important to the present study. Further, there is evidence that multi-faceted Rasch provides better estimates. Alharby (2006) compared the two approaches and two scoring methods, holistic and multiple-trait, within the context of writing performance assessment. He found that in writing assessment utilizing holistic scoring—which is the case with the present study—the Rasch approach resulted in better overall fit. In addition, he compared the estimates achieved using the different measurement and scoring methods to several external measures of writing ability. Multiple regression analysis showed that holistic scores analyzed using multi-faceted Rasch had the highest amount of variance accounted for.

Given the above findings, and given the many studies of performance assessments that have successfully used multi-faceted Rasch measurement, the present study will use the same approach to evaluate the comparability of prompts and the rating quality of raters. In the following sections the chapter provides an overview and description of the Rasch model, its multi-facet extension, and the measures produced by this kind of analysis.

The Rasch Model

In its simplest form, the Rasch approach (Rasch, 1980) attempts to model one underlying attribute at a time by considering the relationship between a person's ability and an item's difficulty. If the data are unidimensional, reflecting one attribute, the relationship will be such that the probability of a correct answer is the difference between a person's ability and an item's difficulty. This can be expressed as follows:

$$P = B_n - D_i$$

where P is the probability of a correct response, B is the ability of a person n, and D the difficulty of an item i. This equation indicates that the more of the underlying attribute a person has, the higher that person's probability of answering the item correctly. Likewise, the more difficult an item is, the lower a person's probability of getting it correctly. In this formulation, item difficulty is independent of but related to person ability. The values of person ability and item difficulty are individually estimated and compared to the data matrix, and the estimates are adjusted repeatedly until the estimate most accurately reflecting the data is arrived at. The natural logarithms of these odds are then calculated, and the measurement scale constructed from these is in log odd units or *logits*. In this transformation, the ordinal relationship between ability and difficulty are now expressed on a common, interval level measurement scale (Bond & Fox, 2007; McNamara, 1996). Thus, there is now information about order and distance, and ability and difficulty can be meaningfully compared.

The final equation for the Rasch model, then, can be expressed in the following manner:

$$P_{ni}(x_{ni} = 1 / B_n, D_i) = \frac{e^{(Bn-Di)}}{1+e^{(Bn-Di)}}$$

where P_{ni} (x_{ni} = 1 | B_n, D_i) is the probability of person n getting the correct answer x = 1 on item i. This probability is equal to the natural log function e raised to the difference between a person's ability and an item's difficulty, divided by one plus the same value. To restate what this equation affords, the Rasch model is "a logistic latent trait model of probabilities which permits items and persons to be analyzed independently, yet still be compared using a common frame of reference" (O'Neill & Lunz, 1996, p. 1).

The Rasch model is often seen as belonging to the item response theory (IRT) family of models. While there are similarities between them, there are actually important differences between them as well, enough for some to view them as conflicting theories (Hambleton, 1989). Rasch and IRT both indeed model test data in an attempt to draw conclusions about person ability and item difficulty. The Rasch model, however, always only focuses on one parameter, item difficulty, to estimate person ability. In IRT, additional parameters can be introduced to account for item discrimination and guessing. The idea behind IRT is to account for the data as well as it can so as to minimize random variation, providing the best fit for the data observed. Where some other data are being considered, however, the best model can well be a different one. Thus, in IRT, the model is made to fit the data. The opposite is true with the Rasch model. Rasch is not concerned so much with the particular data at hand as it is concerned with fundamentally measuring the attribute of interest; thus, one parameter. And to arrive at fundamental measurement, unlike in IRT, the data need to fit the model. Only when that is the case can one claim to have a measurement scale with interval measurement properties (Bond & Fox, 2007).

Estimates and Statistics

To see whether the data indeed fit the model, and indeed measure one attribute at a time, fit statistics can be calculated. It must be remembered that the straight line the Rasch model fits is an idealization, and that few things in reality are completely unidimensional. As well, psychologically multidimensional realities can often exhibit psychologically unidimensional characteristics (Henning, 1992), and for the Rasch model to hold, all that is needed is for one dimension to be sufficiently large. To see if this is indeed the case, fit statistics such as outfit and infit mean square statistics (Wright, 1984; Wright & Masters, 1981) are often calculated and provided by Rasch analysis programs. The outfit mean square statistic individually squares the standardized residuals for every item a person responds to, sums these and divides the total by the number of items the person encountered. It is more likely to be influenced by outlying observations. The infit mean square differs from outfit in that each item is weighted according to its variance, which is larger for on-target observations and smaller for extreme observations. It is more inlier sensitive and more sensitive to violations of unidimensionality, and for those reasons, often more useful than outfit as a fit statistic (Henning, 1992).

Fit statistics are reported both as mean squares and in standardized forms (e.g., *t* or *z*). Mean squares have the form of a ratio scale and an expected value of +1, and a range of 0 to positive infinity. Mean square values greater than one indicate more variation between observed and predicted response patterns than would be expected if the data and model matched perfectly. Values less than one show less variation than would be expected, or overfit. There are no hard and fast rules on what constitute acceptable fit statistics, and what is acceptable can depend on the type of test being analyzed (Wright, Linacre, Gustafsson, & Martin-Loff, 1994). Wright, et al. (1994) suggest 0.4 to 1.2 for judged exams where agreement is encouraged, McNamara (1996) proposed 0.75 to 1.3 as a rule of thumb, and Linacre (2002) offers that 0.5 to 1.5 is acceptable. As for *t* and *z* scores, which vary around a mean of 0 and can be positive or negative, two standard deviations are considered the threshold, with scores above +1.96 reflecting significant misfit and scores below -1.96 reflecting significant overfit.

As mentioned earlier, all estimates are expressed on a common scale in terms of logits. Where a person's ability value and an item's difficulty value are the same, it means that the person has a 50% chance of getting that item correct. A standard error is also provided for each person and item, showing the reliability of their measures. The more observations, the lower the standard errors tend to be. Conversely, the fewer observations, as is the case at the extreme ends of the scale, the higher the standard errors tend to be. In addition to the logit measure, software programs also often report in raw score units the score actually observed and the score expected by the model. Finally, Rasch analysis also provides a separation index and the reliability of that separation. A high separation index and reliability close to 1 would indicate that elements within a facet are indeed being separated into different levels (McNamara, 1996). In most cases, e.g., for persons, this would be a good thing, as it would indicate that the test is able to distinguish test takers with different levels of ability. On the other hand, in the case of raters, it might not be desirable that their levels of severity differ. A chi-square statistic is also provided with the separation index, to see whether observed differences are due to chance; a significant chi-square statistic would indicate that differences are real and not due to chance.

The Multi-Facet Rasch Model

Multi-faceted Rasch measurement (Linacre, 1989) is an extension of the Rasch model that allows researchers to study additional facets beyond person ability and item difficulty which may have an effect on the same. Raters, for example, are thought to possibly have an effect on estimates of test-takers' abilities and on prompt difficulty. Facets such as this can be included in the model. The multi-faceted Rasch model can be written out as follows:

$$P_{nikj} = \frac{e^{(Bn-Di-Fk-Cj)}}{1+e^{(Bn-Di-Fk-Cj)}}$$

where B_n is the ability of person n

D_i is the difficulty of prompt i

F_k is the threshold of score k

C_j is the severity of judge j

Alternately, it can also be expressed in the following fashion:

$$\log (P_{nikj} / P_{nikj} - 1) = B_n - D_i - F_k - C_j$$

As in the original Rasch model, the multi-facet model independently estimates all facets and places them on a common interval scale for comparison. Thus, test-taker ability and prompt difficulty can be objectively measured regardless of the severity or leniency of individual raters.

Research such as this one can also be interested in studying interactions or bias, "construct-irrelevant components that result in systematically lower or higher scores for identifiable groups of examinees" (AERA, et al., 1999, p. 76). Within multi-faceted

Rasch, bias is calculated in the following way: first, measures are estimated for every facet following usual procedures. Then, all facets are anchored except the facet of interest, and bias terms are estimated from the residuals left over from the initial analysis. Bias estimates are reported on the same scale as estimates for other facets, as well as with individual standard errors, again allowing for meaningful comparisons. While the main analysis can sometimes identify aberrant facet functioning, through fit statistics, smaller systematic variance might be overwhelmed by larger components of the responses. This additional step in the analysis helps to verify the validity and fairness of a test.

Earlier, it was stated that an adequate approach to measuring prompt and rater effects needed to have several characteristics, and it can be seen that multi-faceted Rasch fulfills each of those requirements. Prompt difficulty and test-taker ability can be independently estimated while also accounting for rater severity. In addition, estimates within Rasch, while reported as a point, are actually functions that vary with other facets. Different elements within a facet can be rank ordered and real distances between them known, and individual standard errors help to indicate how well the ordering is being done. A mechanism is also available for investigating interactions. And finally, all estimates are reported on the same scale, making meaningful comparisons possible.

Procedures and Analyses

The software program FACETS (Linacre, 2006), which operationalizes the multi-facet Rasch model, was used to analyze the study's data. To fit the requirements of the software, and to make the results more easily interpretable, the ratings—which in the original ranged from 53 to 97 (Appendix B)—were converted into a 0 to 9 scale, where 0

= 53 and 9 = 97. Thus, where results are expressed in terms of scale points, one point represents the difference between one scale point and the next higher or lower scale point.

FACETS needed to be run more than once to answer the questions posed by this study. The command file for the main run is given in Appendix G. In the command file, I specified a model with seven facets: examinee, gender, first language background, language proficiency, perceived prompt difficulty, prompt, and rater (line 7). Gender, first language background, and language proficiency are interaction variables that are the subject of the third research question, while perceived prompt difficulty is the subject of the sixth research question. Language proficiency—in this case, the average of test takers' listening and GVR scores, recoded into ten point bands (i.e., 50-59, 60-69…)—and perceived prompt difficulty are both dummy variables included only for the purpose of examining interactions, and as such were anchored to zero (the average) so that they would not affect the estimates.

This FACETS analysis produced an overall scale—expressed in terms of logits—and all the facets were placed onto this scale, making meaningful comparisons between them possible. So as to provide a frame of reference for subsequent discussions, the overall scale will be presented first in the results.

Prompt-Related Research Questions

Question 1

The study's first research question deals with the degree to which prompts are comparable in difficulty and whether or not there is a prompt effect in the MELAB writing test. The comparability of prompts was evaluated, prima facie, by looking at the prompt measurement report, in particular, by examining the separation index, the reliability of the index, and the results of the chi-square test. These statistics provide global measures for the set of 60 prompts as a whole. The chi-square test tests the null hypothesis that the prompts are equal in difficulty. The separation index, assuming its reliability is high, indicates how many different levels of difficulty the data can be divided into. I examined these measures to see if the prompts were all comparable in difficulty.

The comparability of prompts was also evaluated by considering the difficulty of each prompt. This information is available in two forms. The first are prompt difficulty parameters, which are expressed in terms of logits, the same unit as all other facets. In addition, there is also a fair measure average for each prompt, expressed in terms of the original scale, (in this case, the 0 to 9 scale). I examined the difference between the prompt difficulty parameters and the fair measure averages of the easiest and the most difficult prompts, and determined the effect of these on scores by comparing their spread to the spread of the scale.

Question 2

Question 2 is: “To what extent can it be shown that there is no prompt effect related to topic domain, rhetorical task, prompt length, task constraint, expected grammatical person of response, or number of tasks?” To answer this question, I took the fair measure averages for all 60 prompts and entered them into SPSS 16.0 for Windows. I then entered each prompt’s codes for the six prompt dimensions being investigated.

(Prompt codes can be found in Appendix C.) Cases where coders chose not to agree, as discussed earlier in this chapter, were treated as missing data. Separate analyses of variance (ANOVA) were then conducted for each of the six prompt dimensions. For each ANOVA, the categories within a dimension were the independent variables, and the fair measure averages were the dependent variables. I examined the results of the F-test and the associated p-value for each ANOVA.

Of the six ANOVAs, one was significant. For that prompt dimension, I tested for homogeneity of variances (Levene's statistic), and as that condition held, I used Tukey's HSD as a post hoc test to see which categories were significantly different from each other. I determined this by looking at the mean difference between each pair of categories within that dimension, and by examining the p-values associated with each pair.

Question 3

The third research question concerns possible interaction effects, in particular, whether particular prompts are differentially difficult for test takers of different genders, language backgrounds, and proficiency level. To answer this question, I specified three separate bias/interaction analyses in FACETS. The command "?, ?B, ?, ?, ?, ?B, ?, R9" (line 22 of the command file in Appendix G), specified that bias/interaction analysis be conducted between the second and sixth facets, which in the case of my data are gender and prompt. The bias/interaction analysis between language and prompt was specified in turn, by the command "?, ?, ?, ?B, ?, ?B, ?, R9". Finally, the command for language background and prompt—"?, 1-59B, ?, ?, ?B, ?, R"—was specified so as to exclude from the analysis test takers whose first languages were classified under the "other" categories.

In the output of the bias/interaction analyses, I examined the results of the chi-square tests, which test the null hypothesis that all the combinations (e.g., of prompt and gender) are equal in difficulty. If the null hypotheses had to be rejected, and interaction effects were indeed present, I looked in the results for appropriately measured significant bias values; that is, those with z-scores higher than |1.96| and infit mean square values within the acceptable range. I then evaluated the difference between observed and expected scale point averages for those combinations, to find out the direction of the bias, whether for or against, and the magnitude of the bias.

Rater-Related Research Questions

Question 4

Question 4 is the general rater-related question: "Consistent with the requirements of test validity, reliability, and fairness, to what extent do raters in a large-scale English language proficiency examination rate appropriately and consistently, and to what extent does the test reflect the absence of a rater effect?" This question involved analysis similar to Question 1. First, I examined the rater measurement report for the separation index, the reliability associated with it, and the results of the chi-square test. These global measures show whether raters' severities are comparable or not, and if not, into how many different levels of severity they can be divided, thus providing one indication of rating quality. I also examined individual raters' severity parameters, again comparing their spread to the overall spread of the scale. I evaluated raters' consistency by looking at their infit mean square statistics, where overfit indicates insufficient variation in ratings, otherwise known as the error of central tendency, and where underfit indicates too much variation in ratings or inconsistency. Finally, in order to investigate the possible error of restriction of range, or the inability to distinguish between levels of ability, I considered the fit statistics, the distribution of test takers on the overall scale, and the rating scale itself. Sufficient variation in ratings would provide one indication that raters are able to distinguish different levels of test taker ability, as would seeing test takers spread across the scale. The range covered by each point of the rating scale also provides an indication whether there are a sufficient number of levels or not.

Question 5

The fifth research question is concerned with the characteristics of raters, and has three components to it. The first seeks to find out whether novice raters rate any differently than experienced raters do. The second is related to the first in seeking to find out whether raters rate in the same manner over time. The third aims to discover whether raters' language background—in particular, if English is not their first language—has an effect on the quality of their ratings.

The first two components involved the same operations. To address these, the general idea was to divide the data into smaller sets of data according to time and do separate FACETS analyses of each. There would thus be different severity estimates and fit statistics for each rater, one set for each period of time, and these could be plotted to see if the way they rated changed over time. In Lumley's (2006) study, he employed a cumulative approach, where the first run involved Time 1, the second run involved Time 1 plus Time 2, and so on and so forth. Following this procedure potentially understates changes in rater severity and fit somewhat. On the other hand, it minimizes the possibility of inaccurate estimates as a result of small n sizes. For the study, I decided to follow his procedure to see whether his results could be replicated.

There were, however, several issues that needed to be dealt with before I could undertake the analysis. One, while my data covered more than four years, raters cycled in and out of the data. Some stopped working for the ELI, while others had periods of time where they did not rate any compositions. Thus, if the data were divided according to time, there would not be the same raters or the same number of raters in each smaller set of the data, changing the frame of reference each time and affecting the accuracy of the estimates. Second, the study's data was matched using an approach that just created connectedness for the entire data. Dividing the data into smaller sets of data would result in possible connection issues.

Given the above issues, I looked for periods of over one year where a good number of raters had a good number of ratings throughout the period, and then limited the analysis to those raters. In addition, the periods of time had to begin at points where new raters started rating. This would give me estimates and statistics for when they were novice raters, enabling me to answer the first component of this question. I was able to identify two periods of time with the desired characteristics (Table 4.7). In each of the time periods identified, at least six raters could be included in the analysis. Taken together, there would also be data regarding the rating behavior of four novice raters, from when they started rating and as they progressively gained more experience.

Because even in the below identified periods there remained gaps in ratings, where a rater either did not read or read only a small number of compositions, I decided to use three-month periods as the unit for time. This was to ensure that the resulting estimates were accurate.

Table 4.7 Time Periods for Longitudinal Analysis of Raters

Time Period	Number of Quarters	Raters (* new raters during time period)
09/2004 to 05/2006	7	R01, R09*, R11, R12, R21, R23*
11/2006 to 01/2008	5	R01, R07*, R09, R10*, R20, R22, R23

A solution was also available for the issue of data connectivity. The main FACETS run for the whole set of data, which is connected, had produced estimates for each of the facets. I thus used these known, stable estimates and anchored the facets—with the exception of rater, of course—to these values in the runs with the smaller data sets. The two periods of time above required running FACETS 12 times, resulting in cumulative quarter-by-quarter severity estimates and fit statistics for the raters involved. I plotted the numbers and analyzed them for patterns and regularities (or irregularities, as might be the case) in their rating behavior.

Analyzing the third component of Question 5 was more straightforward than the first two components. First, I revisited the rater severity estimates and the fit statistics that were analyzed in Question 4 above and located the four raters in the study for whom English was not a first language. I examined whether these raters were on the extreme ends of the severity continuum, and whether their ratings had issues related to fit. I also specified a bias/interaction analysis between raters and language background to look for those instances where raters' and test-takers' first language backgrounds were identical. I then examined the z-scores for significant findings, and the difference between expected and observed scores to evaluate the magnitude of any bias.

Question 6

Question 6 asks: "To what extent can it be shown that raters do not alter their rating behavior depending on perceived differences in prompt difficulty or perceived proficiency-related prompt selection behavior among test takers?" I approached this question by specifying a bias/interaction analysis between raters and prompts. I then took the results—in particular, the difference between observed and expected scores—and matched these to the respective prompts in Forms A and B.

In Form A, raters indicated on a five-point Likert-type scale whether a prompt was, in their opinion, easier or more difficult than the average prompt in the MELAB pool (Appendix E). I conducted ANOVAs for each rater, with the five perceived levels of difficulty as the categories being compared. The results were examined for F-statistics with significant p-values associated to them. This analysis allowed me to answer the first part of Question 6, whether raters adjusted their rating behavior depending on perceived differences in prompt difficulty. In Form B, raters identified prompts they thought lower-level candidates were more likely to choose (Appendix F). The comparison groups, in this case, were prompts lower-level candidates were more likely to choose, and prompts they were not more likely to choose. As there were only two categories, I conducted individual independent samples t-tests for each rater. I then examined the t-statistics and p-values for significant differences in the way raters scored the two categories. This analysis allowed me to answer the second part of Question 6, whether raters adjusted their rating behavior depending on perceived prompt-selection behavior among lower-level candidates.

Limitations of the Study

This study has a few limitations, related primarily to the use of available data. First, the test collected only one writing sample from each person. Coupled with the provision of letting test takers choose between two prompts, comparing prompts became a challenge. Sampling and selection issues were created. Fortunately, a way was found to create connections and make comparisons possible, through the matching of certain test takers who took the exam more than once. The matching procedure was more stringent than in most other studies, but it remains that any findings related to prompt effects are only as good as the rationale behind the matches are accepted. As well, the writing samples collected were generally all of a kind, in the mold of traditional compositions. Other types of writing tasks such as data commentary were not used. For that reason, of the two general threats to validity—construct underrepresentation and construct irrelevant variance (Messick, 1989)—the study can only address the latter, but not the former, limiting its scope somewhat.

Second, the use of available data put a limit on the number of issues that could be investigated and the extent to which they could be investigated. The review revealed prompt and rater characteristics that are not reflected in the data. To give just one example, the literature revealed interesting hints about an interaction between raters' professional backgrounds, rating scale, and test-taker ability. But as raters in this data all share the same work background, that issue could not be investigated. Some other issues, such as rater experience and rater language background, the study could address in some way. But while there are new raters and raters of different language backgrounds in the study, their number is small, limiting generalizability. Also, while the data set is relatively large, it was not collected such that all variables are represented in equal number. This limited some investigations, as some cells became too small if not empty. Having pointed out these limitations, the opposite should not be forgotten: the study is using real-world data and using an appropriate method to address a large number of issues concerning prompt and rater effects, which has been called for often (Connor-Linton, 1995a; Kane, et al., 1999; Lane & Stone, 2006) but not been done before.

Chapter Summary

This chapter described the data and the methods I used to address the research questions regarding prompt and rater effects in writing performance assessment. The data for the study came from the MELAB program, and it was argued that the limitations brought about by the use of extant data were more than made up for by the number of issues that could be addressed by the large set of real-world data. The primary method chosen to answer the questions was the multi-facet extension of the Rasch model. This model was described and shown to have a number of desirable qualities that make it an appropriate method for dealing with the questions. Specific procedures and analyses undertaken to answer the research questions were described in detail. The next chapter presents the results and answers the questions posed by the study, whether systematically varying aspects of writing performance assessments affect the validity, reliability and fairness of these tests.

CHAPTER 5

RESULTS AND DISCUSSION

Performance writing assessments are taken by millions of people each year, and the results of these high-stakes tests affect these people's life chances. Thus, it is vitally important that these tests be valid, reliable, and fair. The purpose of this study is to determine whether that is indeed the case. The main question guiding the study is: **How are the validity, reliability, and fairness of a second language writing performance assessment affected by aspects of the examination that are systematically varied for different test takers?** This question was operationalized into six research questions dealing with prompts and raters, and investigated primarily through the use of multi-facet Rasch analysis. In addressing these questions, I will be able to verify the relationship between a number of prompt- and rater-related factors and the scores that test takers receive, and make an argument about the validity of writing performance assessment.

The specifics of the data and methods used were articulated in the previous chapter. This chapter presents the results of the study and provides a discussion of the same. First, I will present the overall logit scale created by the FACETS analysis, as well as how the MELAB rating scale for writing fits into this scale. This will provide a frame of reference for understanding the parameter values in subsequent parts of the chapter. Then, results of the three questions that have to do with prompts will be presented, followed by results of the three questions that have to do with raters. As has previously been mentioned, the first question in each set of three is a more general question that is each followed by two more specific questions. That is to say, the answer to each general question can change depending on the answers to the more specific questions. Thus, in the presentation and discussion of results, there will be a measure of revisitation and repetition.

The Overall Scale and the Rating Scale

One important advantage of using multi-faceted Rasch analysis is that resulting estimates for each facet and each element are interval measures which are all placed on a common scale. This makes meaningful comparisons between different facets possible. The scale is expressed in terms of log-odds units or logits, with zero as the average. For this study, FACETS generated a scale that ranged from -17 to 17 logits, or approximately 34 logits (Figure 5.1). The breadth of this scale reflects the wide range of test-taker abilities, though the majority of test-takers' ability estimates apparently fall between -7 and 12 logits. Figure 5.1 also shows how the MELAB rating scale maps onto the common logit scale in this model (Column 3). Column 6 shows that, with the exception of the lowest scale point (0, or 53 in the original MELAB scale), the scale points are each the most likely rating for a range of several in the overall scale. Scale point 1 covers the greatest ability range (5.33 logits), while scale point 7 covers the least (2.94 logits), with the average being 3.88 logits. These differences in the logit range covered by each scale point means that the MELAB rating scale is not an interval scale. Thus, for example, going from a 4 to a 5 does not represent the same increase in ability as going from a 5 to a 6 or from a 7 to an 8. On the other hand, the results also show that the 10-point scale is generally appropriate, in that each scale point is in fact used and is the most likely rating for part of the ability range.

Figure 5.1 The Logit Scale and the Rating Scale

```
------------------------------------------------------------------
| Measure |     Test     | Scale |    Most   |  SE   | Range of  |
|         |    Takers    |       | Likely At |       | Scale Pt. |
|----------------------------------------------------------------|
|   17    | .            |  (9)  |           |       |           |
|   16    | .            |       |           |       |           |
|   15    | .            |-------|  14.50    |  .10  |-----------|
|   14    | .            |       |           |       |           |
|   13    | .            |   8   |           |       |    3.87   |
|   12    | *.           |       |           |       |           |
|   11    | *.           |-------|  10.63    |  .05  |-----------|
|   10    | *.           |       |           |       |           |
|    9    | **.          |   7   |           |       |    2.94   |
|    8    | ****.        |-------|   7.69    |  .03  |-----------|
|    7    | *****.       |       |           |       |           |
|    6    | *******.     |   6   |           |       |    3.18   |
|    5    | ********.    |       |           |       |           |
|    4    | *********.   |-------|   4.51    |  .02  |-----------|
|    3    | *********.   |   5   |           |       |    3.57   |
|    2    | *********.   |       |           |       |           |
|    1    | *********.   |-------|   0.94    |  .02  |-----------|
|    0    | ********.    |       |           |       |           |
|   -1    | *******.     |   4   |           |       |    4.24   |
|   -2    | *****.       |       |           |       |           |
|   -3    | ****.        |-------|  -3.30    |  .03  |-----------|
|   -4    | ***.         |       |           |       |           |
|   -5    | **.          |   3   |           |       |    3.85   |
|   -6    | *.           |       |           |       |           |
|   -7    | *.           |-------|  -7.15    |  .05  |-----------|
|   -8    | .            |       |           |       |           |
|   -9    | .            |   2   |           |       |    4.09   |
|  -10    | .            |       |           |       |           |
|  -11    | .            |-------| -11.24    |  .11  |-----------|
|  -12    | .            |       |           |       |           |
|  -13    | .            |       |           |       |           |
|  -14    | .            |   1   |           |       |    5.33   |
|  -15    | .            |       |           |       |           |
|  -16    | .            |  (0)  | -16.57    |  .36  |           |
|----------------------------------------------------------------|
|         | * = 126      |       |           |       | Ave.=3.88 |
------------------------------------------------------------------
```

Prompt-Related Questions

Prompt Comparability

In this section I begin to answer the question: **Consistent with the requirements of test validity, reliability, and fairness, to what extent are the writing prompts in a large-scale English language proficiency examination comparable in difficulty, and to what extent does the test reflect the absence of a prompt effect?**

The question was answered, initially, by considering the separation index, the reliability of the index, and the chi-square statistics. The separation index is the ratio between the adjusted standard deviation (Adj. SD) of the prompt measures and the root mean square error (RMSE), providing an indication of how much greater the observed variance is over the error. The reliability statistic provides an indication of how reliably prompts are being separated into different difficulty levels. In the present case, the lower the separation index and the lower the reliability statistic, the better, as it would indicate that the prompts cannot be divided into different levels of difficulty. For its part, the fixed chi-square tests the likelihood that all prompts are equal in difficulty. A significant finding would mean that prompts are not in fact equally difficult.

The above statistics provided measures of the prompts taken together. In addition, FACETS produced difficulty parameters and fit statistics for each prompt. Regarding prompt difficulty parameters, zero indicates a prompt of average difficulty. Parameters with greater negative values represent easier prompts, and parameters with greater positive values represent more difficult prompts. FACETS provided two sets of fit statistics—the infit mean square and the outfit mean square—which indicate how consistently test takers responded to the prompt. Of the two, the infit mean square is generally considered more useful because it is weighted to favor on-target observations that are more accurately measured (Henning, 1992). Infit mean square values of between 0.4 and 1.5 are generally considered to be acceptable (Linacre, 2002; Wright, et al., 1994). In addition, FACETS also provided a fair measure average for each prompt, allowing one to evaluate prompt difficulty in terms of the original scale.

Results

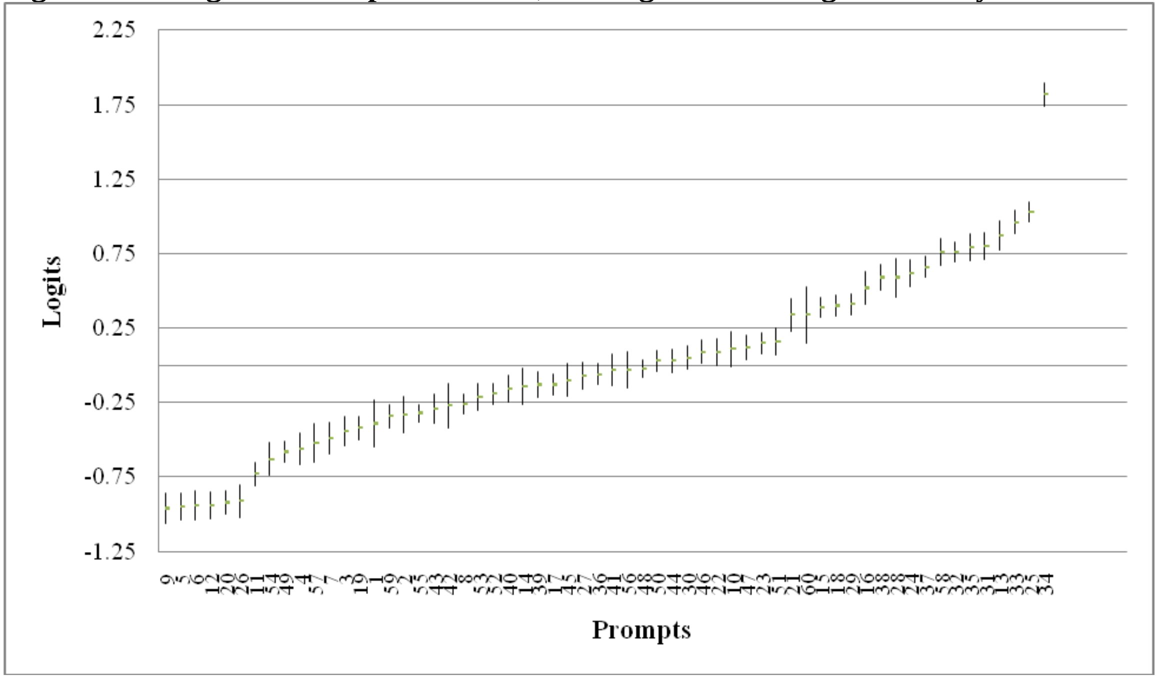
Table 5.1 provides the difficulty parameters for the prompts in this data, arranged in order of difficulty from the easiest to the most difficult. The group level statistics are at the bottom of the table. The separation index for this set of prompts was 5.85, with a reliability of .97, indicating that the prompts can reliably be separated into at least five different levels of difficulty. The fixed chi-square test had a p-value of .00; that is to say, the null hypothesis that the prompts are equal in difficulty must be rejected. The prompts ranged in difficulty from -0.96 to 1.82, or a range of 2.78 logits. In terms of the original scale, the fair average score for the most difficult prompt was 4.36, and 5.13 for the easiest prompt. All the prompts showed acceptable infit statistics, meaning they have been consistently measured. The prompt difficulty measures, accounting for standard error, are shown graphically in Figure 5.2. From the graph, it can clearly be seen that Prompt 34 is a clear outlier. The six easiest prompts also appear to form one level of difficulty of their own.

Table 5.1 Prompt Measurement Report

```
-----------------------------------------------------
|Prompt|   n   Obsvd  Fair |Measure S.E.|  Infit    |
|      |        Ave.  Ave. |            | MnSq ZStd |
|---------------------------------------------------|
|   9  |  321   5.3   5.13 | -.96   .10 |  .8   -3  |
|   5  |  403   5.2   5.12 | -.95   .09 |  .9   -1  |
|   6  |  359   5.3   5.12 | -.94   .10 |  .9   -1  |
|  12  |  475   5.1   5.12 | -.94   .09 |  .9    0  |
|  20  |  546   5.3   5.12 | -.92   .08 |  .8   -3  |
|  26  |  277   5.3   5.11 | -.91   .11 | 1.0    0  |
|  11  |  620   5.1   5.07 | -.73   .08 |  .7   -4  |
|  54  |  326   5.0   5.04 | -.63   .11 |  .7   -4  |
|  49  |  682   5.0   5.03 | -.58   .07 |  .8   -4  |
|   4  |  333   4.9   5.02 | -.56   .11 |  .7   -4  |
|  57  |  201   5.4   5.01 | -.52   .13 |  .9   -1  |
|   7  |  292   5.2   5.01 | -.49   .11 |  .9   -1  |
|   3  |  343   4.9   4.99 | -.44   .10 |  .7   -3  |
|  19  |  601   5.0   4.99 | -.42   .08 |  .7   -4  |
|   1  |  149   5.0   4.98 | -.39   .16 |  .8   -1  |
|  59  |  516   5.0   4.97 | -.34   .08 |  .9   -1  |
|   2  |  243   5.0   4.96 | -.33   .12 | 1.0    0  |
|  55  | 1002   4.9   4.96 | -.32   .06 |  .9   -2  |
|  43  |  330   4.8   4.95 | -.29   .10 | 1.2    2  |
|  42  |  148   5.2   4.95 | -.27   .15 | 1.0    0  |
|   8  |  729   4.9   4.95 | -.26   .07 | 1.0    0  |
|  53  |  495   4.9   4.93 | -.21   .09 |  .6   -6  |
|  52  |  846   4.9   4.93 | -.19   .07 |  .9   -2  |
|  40  |  427   4.8   4.92 | -.16   .09 |  .8   -3  |
|  14  |  265   5.0   4.91 | -.14   .12 | 1.0    0  |
|  39  |  459   4.9   4.91 | -.13   .09 |  .9   -1  |
|  17  |  691   5.0   4.91 | -.13   .07 |  .8   -4  |
|  45  |  318   5.0   4.90 | -.10   .11 | 1.0    0  |
|  27  |  438   4.8   4.90 | -.07   .09 |  .9   -1  |
|  36  |  812   4.8   4.89 | -.06   .07 |  .7   -5  |
|  41  |  302   4.7   4.89 | -.03   .11 |  .7   -4  |
|  56  |  260   4.8   4.89 | -.03   .12 |  .7   -3  |
|  48  |  975   4.9   4.88 | -.02   .06 |  .8   -4  |
|  50  |  797   4.9   4.87 |  .03   .07 |  .8   -4  |
-----------------------------------------------------
```

```
-----------------------------------------------------
|Prompt|   n   Obsvd  Fair |Measure S.E.|  Infit    |
|      |        Ave.  Ave. |            | MnSq ZStd |
|---------------------------------------------------|
|  44  |  623   4.8   4.87 |  .03   .08 |  .8   -4  |
|  30  |  518   4.9   4.86 |  .05   .08 |  .9   -1  |
|  46  |  578   4.8   4.86 |  .09   .08 |  .8   -4  |
|  22  |  493   4.8   4.85 |  .09   .09 | 1.0    0  |
|  10  |  264   4.8   4.85 |  .11   .12 |  .9    0  |
|  47  |  529   4.7   4.85 |  .12   .08 |  .8   -3  |
|  23  |  695   4.8   4.84 |  .15   .07 | 1.0    0  |
|  51  |  427   4.8   4.84 |  .16   .09 | 1.0    0  |
|  21  |  330   4.7   4.79 |  .34   .11 | 1.0    0  |
|  60  |  105   5.2   4.79 |  .34   .19 | 1.2    1  |
|  15  |  732   4.6   4.77 |  .39   .07 |  .8   -3  |
|  18  |  833   4.6   4.77 |  .40   .07 |  .9   -1  |
|  29  |  700   4.8   4.77 |  .41   .07 |  .9   -2  |
|  16  |  328   4.7   4.74 |  .52   .11 |  .6   -5  |
|  38  |  508   4.6   4.72 |  .59   .09 |  .9    0  |
|  28  |  232   4.6   4.72 |  .59   .13 |  .8   -1  |
|  24  |  486   4.8   4.71 |  .62   .09 | 1.0    0  |
|  37  |  764   4.5   4.70 |  .66   .07 |  .8   -5  |
|  58  |  448   4.5   4.67 |  .76   .09 |  .9   -1  |
|  32  |  679   4.4   4.67 |  .76   .07 |  .9   -2  |
|  35  |  453   4.5   4.66 |  .79   .09 |  .7   -4  |
|  31  |  433   4.6   4.66 |  .80   .09 | 1.0    0  |
|  13  |  404   4.5   4.64 |  .87   .10 |  .8   -3  |
|  33  |  633   4.4   4.61 |  .96   .08 | 1.0    0  |
|  25  |  863   4.4   4.59 | 1.03   .07 |  .7   -5  |
|  34  |  620   4.0   4.36 | 1.82   .08 |  .8   -2  |
|---------------------------------------------------|
| Mean |  494.3 4.9   4.87 |  .00   .09 |  .9  -2.4 |
| S.D. |  211.8  .3    .15 |  .57   .02 |  .1   2.1 |
|---------------------------------------------------|
|  RMSE (Model)  .10             Adj S.D.      .56  |
|  Separation   5.87             Reliability   .97  |
|  Fixed chi-square: 2674.0      d.f.:          59  |
|                                significance: .00  |
-----------------------------------------------------
```

Figure 5.2 Ranges of Prompt Estimates, Arranged According to Severity



Discussion

The results indicate that these prompts are not comparable in difficulty. The chi-square test, the separation index, and the reliability statistic all lead to this conclusion. In some way, this is not unexpected, as it is difficult to imagine such a large number of prompts all being entirely equal in difficulty. The real question is whether these significant differences are also meaningful ones as well.

An examination of the prompt difficulty parameters immediately shows that the most difficult prompt, Prompt 34, has an estimate much higher than the rest, more than three standard deviations from the mean. The difficulty parameter of this prompt, accounting for standard error, is somewhere in the range of 1.74 and 2.00, whereas the range for the next most difficult prompt, Prompt 25, is between 0.96 and 1.10. In Figure

5.2, it is clearly seen that there is no overlap between the possible true parameter estimates for these two prompts, and thus they can unambiguously be separated into different difficulty levels. Thus, if just one outlier prompt were removed, the number of levels into which the prompts can be divided would immediately be reduced from five to four. In terms of logits, the range between the easiest and most difficult prompt would be reduced by almost a third from 2.78 to 1.99. If just two more prompts were excluded—say, Prompts 25 and 33—the range between the easiest and most difficult prompts would be further reduced to just 1.83 logits.

Assuming that three of the prompts were excluded, what is the practical effect of the easiest and most difficult prompt differing by 1.83 logits? Rasch, as I have mentioned, makes meaningful comparisons between different facets possible. It can be recalled that the rating scale has also been expressed in terms of the logit scale, and that the average range covered by each scale point is 3.88 logits. Given that, on average, an advantage of 1.94 logits (50% of a scale point) would be necessary for one to get rounded off to the next higher score. Thus, if just three prompts were excluded from the pool, even if the remaining prompts represent four different levels of difficulty, on average, the difference between the easiest and most difficult prompt, 1.83, would have no practical effect on the score a person receives.

The above discussion can be restated in terms of the original scale. Including all 60 prompts, the difference between the easiest and the most difficult prompt is 5.13 - 4.36 = 0.77 points, or about three-quarters of a scale point. However, if the three prompts were to be excluded, the difference between the remaining easiest and most difficult prompt would be 0.5—or at just the halfway point between scale points. Reducing the pool of prompts to 57 would, on average, ensure that scores are not unduly affected because of prompt assignment.

That is, of course, only on average. For example, analysis of the scale in Chapter 2 indicated that the decision point is between scale points 4 and 5. Scale point 4 is wider than the average, spanning a logit range of 4.24. Thus, at the critical decision point, prompt difficulty would have to differ by 2.14 logits to have an effect. On the other hand, scale point 7 only covers a range of 2.94 logits, and differences in prompt difficulty would be more likely to have an effect on actual scores at that scale point. To ensure that there is no prompt-related effect in the test at any point along the scale, the difference between the easiest and most difficult prompt would have to be no larger than 1.47 logits. Approximately 14 of the easiest and most difficult prompts would need to be removed from the pool for this to happen.

To summarize, to the question of prompt comparability and prompt effects, it would appear that the prompts do indeed differ in difficulty and that there can be situations where a prompt effect might exist. However, the possibility of prompt effects is created mainly by a few outlier prompts, and the exclusion of these prompts would be sufficient to bring about a state of affairs where prompt differences have no practical effect on test scores.

Prompt Dimensions

The previous section showed that differences in prompt difficulty do exist. It can be asked whether these differences are random, or if there are particular characteristics and qualities of prompts that make some of them systematically more difficult than others. The research question being considered in this section aims to help answer that by investigating several prompt dimensions that have been identified in the literature: **To what extent can it be shown that there is no prompt effect related to topic domain, rhetorical task, prompt length, task constraint, expected grammatical person of response, or number of tasks?**

This question was answered first, by showing the fair averages for each category within each prompt dimension. Whether any differences are significant was determined by conducting a series of ANOVAs, one for each of the prompt dimensions, with the FACETS fair average measures for the prompts as the dependent variable. Significance was set at $p > .05$. I examined the results of the F-test and their associated p-values to determine significant differences, and conducted post hoc tests where appropriate.

Results

First, Table 5.2 shows the average fair measure scores for different categories within each of the six prompt dimensions, arranged from the easiest to the most difficult. It can be seen that the largest spread between categories can be found within topic domain, about 0.15 of a scale point difference between prompts on education topics and prompts on social topics. For rhetorical task and prompt length, the spread was approximately 0.12 and 0.11, respectively. The spread was less than 0.05 for the remaining three dimensions.

Table 5.2 Fair Averages for Categories within Prompt Dimensions

Topic Domain			Rhetorical Task			Prompt Length		
	n	Fair Ave.		n	Fair Ave.		n	Fair Ave.
Education	6	4.98	Expository	30	4.90	2 sentences	14	4.92
Business	10	4.97	Argumentative	22	4.86	1 sentence	2	4.89
Personal	12	4.86	Narrative	5	4.78	3 sentences	20	4.87
Social	29	4.83				4 sentences	20	4.86
						5 sentences	4	4.81

Task Constraint			Grammatical Person			Number of Tasks		
	n	Fair Ave.		n	Fair Ave.		n	Fair Ave.
Unconstrained	12	4.88	Third Person	32	4.87	1 task	8	4.90
Constrained	40	4.87	First Person	25	4.87	3 tasks	21	4.89
						4 tasks	6	4.87
						2 tasks	22	4.86

Whether the above differences are significant or not can be determined by examining the results of the ANOVAs, which are reported in Table 5.3. Of the six prompt dimensions tested, only topic domain showed significant differences, F(3,53) = 3.858, p = .025. Differences in all other dimensions failed to reach statistical significance.

Table 5.3 Prompt Dimensions Analyses of Variance

	df		F	Sig.
	Between Group	Within Group		
Topic Domain	3	53	3.386	.025*
Rhetorical Task	2	54	1.406	.254
Prompt Length	4	55	0.516	.724
Task Constraint	1	50	0.014	.905
Grammatical Person	1	55	0.017	.897
Number of Tasks	3	53	0.120	.948

For topic domain, a test for equality of variance (Levene's statistic) showed that the assumption of equal variances is valid. Thus, a post-hoc test using Tukey's HSD was appropriate and was conducted to see where the significant difference or differences resided. The post-hoc test, contrary to the ANOVA, did not show any significant differences among the different topic domains (Table 5.4). However, an inspection of the p-values indicated that the difference between business prompts and social prompts, 0.14 of a scale point, was the one closest to significance. The difference between education and social prompts was also marginally close to being significant.

Table 5.4 Mean Differences and p-values for Post-Hoc Test

Col–Row (Sig.)	Business	Education	Personal	Social
Business	.000	-.013 (.998)	.104 (.362)	.140 (.057)
Education		.000	.117 (.394)	.153 (.106)
Personal			.000	.036 (.888)
Social				.000

Discussion

The above analysis considered six dimensions of prompts, and whether certain categories within each created systematic differences in scores and in prompt difficulty. Of the six, a significant difference might or might not exist only within one dimension. The ANOVA and the post-hoc test disagreed on whether the difference found between prompts on business topics and prompts on social topics was significant. In addition, there were relatively few prompts coded under the education domain, and this may or may not have affected the lack of significant findings between education prompts and social prompts, which have the largest mean difference between them.

Significance aside, the difference between the two topic domains that may or may not be significant amounted to 0.14 of a scale point—not likely to make a difference in the final score in most situations. (It might also be worth noting that the outlier prompt identified earlier, Prompt 34, as well as 8 of the 12 most difficult prompts, relate to the social domain. Thus, the same process of excluding a few outlier prompts can likely take care of this problem without much difficulty.) The relatively small differences in scores obtained means that, no matter the topic domain assigned, test takers are generally able to produce compositions of comparable quality. This provides evidence for one aspect of the test's design indeed being the case, the use of topics presumed to be familiar to all test takers.

The general lack of findings here conforms to much of the literature. It has been noted, for example, that expected grammatical person of response is not usually very salient to test takers (Greenberg, 1981), while fulfillment of tasks given in a prompt is not usually an important consideration for raters (Connor & Carrell, 1993). Besides, tasks can differ in the length and complexity of response required, from one word (e.g., "Do you agree or disagree?") to several paragraphs (e.g., "Discuss.") Because of this, number of tasks just does not capture the complexity or difficulty of a prompt very well. For its part, task constraint was intended to capture the number of ways a test taker could respond to a prompt. It appears that having different ways of responding to a prompt was not all that important, given that (1) one only really needs to give one response, (2) the prompts are apparently generally accessible anyway, and if one prompt was not accessible, (3) test takers could choose to write on the other prompt. There was an apparent pattern where length of prompts is concerned. The only category that was out of order was one sentence prompts. There were, however, only two one-sentence prompts. That category aside, there is an inverse relationship between prompt length and fair average score; the longer the prompt, the lower the average score. However, this relationship was not significant. Reading a longer prompt might take somewhat more time, but not all that much (cf. Polio & Glew, 1996).

One dimension that is much discussed in the literature is the rhetorical task required by the prompt. As was the case in many other studies (Hamp-Lyons & Mathias, 1994; Quellmalz, Capell, & Chou, 1982; Wiseman, 2009), prompts calling for a narrative response had the lowest fair average. Argumentative prompts, however, did not have the highest fair averages; expository prompts did. Again, though, it must be noted that these differences were not significant. Like experts in other studies, the raters in this study appear to think narrative prompts are easier to respond to (see results to Question 6); whether this perception caused raters to adjust their rating behavior will be seen later in the chapter.

The one dimension that yielded possibly significant differences was topic domain. Interestingly, when asked what factors they considered in choosing prompts, test takers overwhelmingly cited background knowledge and topic familiarity (Polio & Glew, 1996; Powers & Fowles, 1998). Their intuition is apparently correct as, in this test at least, topic domain seems to be the only dimension of prompts that might have an effect on scores at all.

Prompts and Test-Taker Characteristics

The third research question investigates the relationship between prompts and different test-taker characteristics: **To what extent are writing prompts not differentially difficult for test takers of different genders, language backgrounds, and proficiency level?** To investigate this question, I examined the results of the bias/interaction analysis between the prompts and each of the three test-taker characteristics. The results of the chi-square test indicate whether the null hypothesis, that there is no differential effect, should be rejected. If differential effects do exist, estimates for individual combinations of prompts and test-taker characteristics can indicate at a more fine-grained level where the bias exists; in particular, the z-scores, the infit mean square statistic, and the difference between observed and expected scores for each combination of prompt and test-taker characteristic.

Results

Results of the bias/interaction analysis between prompt and gender, language background, and test-taker proficiency level are given in Tables 5.5, 5.6, and 5.7, respectively. Provided in the tables are the global measures, as well as individual interaction measures that are significant (|z-score| > 1.96). It can be seen that for all three analyses, the significance of the chi-square tests was 1.00. That is, the null hypothesis that there is no differential effect should not be rejected. In all three analyses, the average difference between observed score and expected score for the different interaction terms was 0.01 of a scale point. In the case of prompt and language background, however, three combinations yielded significant results, two involving Sinhalese speakers, and one involving Spanish speakers. The significant results included bias in both directions, for and against indicated native speaker groups.

Table 5.5 Bias/Interaction Analysis: Prompt and Gender

```
----------------------------------------------------------------------
|      Prompt x        | Obs-Exp |  Bias+   Model         |Infit Outfit|
|       Gender         | Average | Measure   S.E.  Z-Score| MnSq  MnSq |
|--------------------------------------------------------------------|
| Mean (Count: 120)    |   .01   |   -.04    .14    -.29  |   .9    .8 |
| S.D.                 |   .01   |    .03    .04     .20  |   .2    .2 |
|--------------------------------------------------------------------|
| Fixed chi-square: 15.4  d.f.: 120  significance: 1.00              |
----------------------------------------------------------------------
```

Table 5.6 Bias/Interaction Analysis: Prompt and Language Background

```
----------------------------------------------------------------------
|       Prompt x       | Obs-Exp |  Bias+   Model         |Infit Outfit|
| Language Background  | Average | Measure   S.E.  Z-Score| MnSq  MnSq |
|--------------------------------------------------------------------|
| 13 x Sinhalese (2)   |  -.83   |   3.56    1.33    2.69 |   .9    .9 |
| 43 x Sinhalese (2)   |   .84   |  -3.03    1.26   -2.41 |   .7    .7 |
| 60 x Spanish   (4)   |  -.55   |   2.44    1.02    2.40 |  2.0   2.1 |
|--------------------------------------------------------------------|
| Mean (Count: 2103)   |   .01   |   -.04    .84    -.06  |   .7    .7 |
| S.D.                 |   .05   |    .19    .40     .21  |   .8    .8 |
|--------------------------------------------------------------------|
| Fixed chi-square: 102.6  d.f.: 2103  significance: 1.00            |
----------------------------------------------------------------------
```

Table 5.7 Bias/Interaction Analysis: Prompt and Proficiency Level

```
----------------------------------------------------------------------
|       Prompt x       | Obs-Exp |  Bias+   Model         |Infit Outfit|
|  Proficiency Level   | Average | Measure   S.E.  Z-Score| MnSq  MnSq |
|--------------------------------------------------------------------|
| Mean (Count: 358)    |   .01   |   -.02    .31    -.12  |   .8    .8 |
| S.D.                 |   .02   |    .06    .21     .26  |   .4    .4 |
|--------------------------------------------------------------------|
| Fixed chi-square: 29.1  d.f.: 358  significance: 1.00              |
----------------------------------------------------------------------
```

Discussion

The results of the bias/interaction analysis for prompt and gender and for prompt and proficiency level are straightforward. They unequivocally show that prompts are not differentially difficult for test takers according to those two characteristics. The results for prompt and language proficiency, however, require some further discussion. In that analysis, the chi-square test indicates that, overall, bias does not exist. However, in the results for individual combination, three out of 2,103 bias terms had z-scores that were significant. The bias term for the combination of Spanish and Prompt 60 had high infit and outfit measures associated with it, indicating that the observations do not fit the model very well and that other things were affecting the estimate. As such, this particular finding should be discounted. The two "meaningfully" significant bias terms both involve test takers who speak Sinhalese as a first language. Prompt 13 was more difficult than expected, according to the analysis, as indicated by the negative observed-minus-expected value, whereas Prompt 43 was easier than expected. These measurements, however, are each based on two ratings; because compositions are always double rated, that means one test taker each.

There are two ways of interpreting the findings. One way of interpreting them would be that the two test-takers' abilities are typical of their language group, and that the prompts are indeed easier and more difficult, respectively, for Sinhalese speakers. The biases would then apply to all other Sinhalese test-takers in the study. The other way of interpreting the findings would be that the two test-takers' abilities are not typical of their language group, but as the bias/interaction analysis was conducted based on the measure for their group rather than on their individual measures, apparently significant but spurious results were found. It is difficult to think that the first interpretation is the correct one. If there is something about prompts that makes them biased, what accounts for the observed biases? Why are the observed biases in different directions? And why are the biases not reflected in any of the other 58 prompts? Or among those whose language background and culture are similar to the Sinhalese? The second interpretation is more plausible. Given the results of the chi-square test, given the absence of significant findings in over 2,000 bias terms, and given that the only two significant findings are each based on n-sizes of one, it is more likely that the significant findings are artifacts of estimation based on inadequate samples, and are in fact false. Thus, it would be appropriate to conclude that where prompt and language background is concerned, as with the other two background factors, there is in fact no interaction effect.

In the literature, an interaction is sometimes observed between prompt and the three test-taker background characteristics discussed here (e.g., Breland, et al., 2004; Broer, et al., 2005; Gabrielson, et al., 1995; Lee, et al., 2004). Significant findings usually involved only a few prompts from within their respective pools, and effect sizes were usually small. (On the other hand, there are also studies that show no interaction effect, e.g., Park, 2006). In general, there are a few differences between those studies and the current one. First, those studies were generally based on stronger assumptions, in that all test takers were matched according to an English language-ability variable. The current study matched a smaller number of test takers under more stringent matching conditions, allowing other test-takers' abilities to be statistically estimated rather than a priori assumed. Second, the other studies' interaction analyses were based on residuals after accounting for ability and the variable of interest. The current study's bias/interaction analyses were conducted on residuals after multiple explanatory variables had been accounted for in the main estimation. There is thus presumably less unexplained variance left to explain. Finally, the other studies employed logistic regression, and as a result of making stronger assumptions could compare test-taker background characteristics directly. The current study employed multi-faceted Rasch, and as people cannot belong to more than one category for each background characteristics, interaction analysis was done indirectly. That is, the comparison is between the expected score and observed score of, say, a male test taker on that prompt, rather than a comparison between the scores of male and female test takers. Since the difference between observed and expected score of male and female test takers are not added up, the bias presumably appears smaller, and perhaps for that reason goes undetected. Of the three differences between this study and other studies, the first two are reasons for thinking the results of the present study are more dependable, whereas the third is a reason for thinking that the present study underestimated and failed to detect real differences. In any case, on the whole, the present study agrees with others in concluding that much of the differences observed, when they are observed, are not examples of item bias but rather of item impact (Clauser & Mazor, 1998; Penfield & Lam, 2000; Zumbo, 1999). That is, differential probabilities of success are attributable to actual differences in the ability of interest.

Section Summary

The first three research questions dealt with the possibility of a prompt effect. That is, that some prompts, because of some feature, are more difficult than others prompts, whether in general or for particular groups of test takers, resulting in scores that are not valid, reliable, or fair. The results of Question 1 indicate that the prompts are on the whole comparable in difficulty, except for a few prompts that are more difficult to get a high score on. A possible reason why those prompts are more difficult was seen in the results of Question 2. Prompts belonging to the social domain tended to be more difficult, but statistically so perhaps only when compared with the easiest domain, business. Test-takers' gender, language background, and proficiency level do not appear to cause differential prompt difficulty, according to the results of Question 3.

Having answered the more specific prompt-related questions, a more definitive answer can now be given for the more general, overall prompt-related question: Consistent with the requirements of test validity, reliability, and fairness, to what extent are the writing prompts in a large-scale English language proficiency examination comparable in difficulty, and to what extent does the test reflect the absence of a prompt effect? The study showed that while the prompts are not all comparable in difficulty, the differences were such that they would not generally have an effect on final scores, and that an argument can therefore be made that there is no threat to the validity, reliability, or fairness of this test due to a prompt effect.

Rater-Related Questions

Rating Quality

Raters are central to the enterprise of performance assessments, as they are the ones who actually do the assessing. Thus, it is important to know: **Consistent with the requirements of test validity, reliability, and fairness, to what extent do raters in a large-scale English language proficiency examination rate appropriately and consistently, and to what extent does the test reflect the absence of a rater effect?**

The question's interest is in the general rating tendencies of individual raters, whether or not they have problems of severity, central tendency, or restriction of range. As well, it is also interested in the collective performance of raters, whether together the ratings they give result in scores that are valid, reliable, and fair. Similar to the first research question, this question was answered by evaluating the prompt measurement report; in particular, by examining the separation index, the reliability of that index, and the chi-square test statistics, and by examining the severity estimates and fit statistics for individual raters. In addition, in order to evaluate possible restriction of range, the overall scale and the rating scale were also considered.

Results

The rater measurement report is given in Table 5.8, and is arranged according to rater severity estimates, from the most lenient to the most severe. Global statistics are found at the bottom of the table. Considering the raters as a group, the significant chi-square test ($p < .00$) indicated that the raters do not exhibit equal degrees of severity. The separation index (2.86) suggested that they can be divided into almost three different levels. The raters' severity estimates ranged from -0.89 to 1.33, or a range of 2.22 logits. Infit statistics generally fell within the acceptable range of 0.4 to 1.5, with the exception of raters R04 and R13, who had infit mean squares of 2.8 and 1.7, respectively .

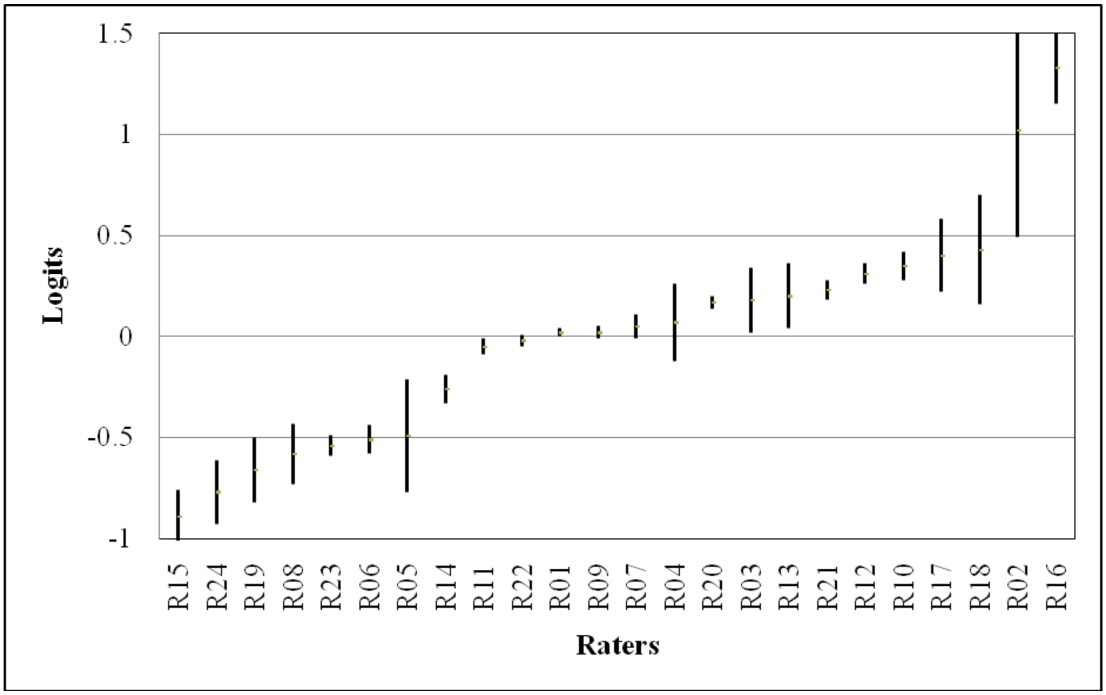
Table 5.8 Rater Measurement Report

```
-------------------------------------------------------
| Rater | Count |          Model |  Infit     Outfit   |
|       |       | Measure  S.E. |MnSq ZStd  MnSq ZStd |
|-----------------------------------------------------|
|  R15  |  232  |   -.89   .13  |  .9  -1     .9  -1  |
|  R24  |  139  |   -.77   .16  |  .8  -1     .8  -2  |
|  R19  |  133  |   -.66   .16  |  .9   0     .9   0  |
|  R08  |  158  |   -.58   .15  | 1.4   2    1.3   2  |
|  R23  | 1220  |   -.54   .05  |  .9  -3     .8  -4  |
|  R06  |  760  |   -.51   .07  | 1.0   0     .9  -1  |
|  R05  |   45  |   -.49   .28  |  .9   0     .8   0  |
|  R14  |  685  |   -.26   .07  |  .7  -5     .7  -5  |
|  R11  | 2379  |   -.05   .04  |  .8  -6     .8  -7  |
|  R22  | 3231  |   -.02   .03  |  .6  -9     .6  -9  |
|  R01  | 6563  |    .02   .02  | 1.0   0     .9  -3  |
|  R09  | 5124  |    .02   .03  |  .7  -9     .7  -9  |
|  R07  |  937  |    .05   .06  | 1.0   0    1.0   0  |
|  R04  |   97  |    .07   .19  | 2.8   8    2.6   7  |
|  R20  | 3351  |    .17   .03  |  .9  -6     .8  -7  |
|  R03  |  138  |    .18   .16  | 1.3   2    1.3   2  |
|  R13  |  145  |    .20   .16  | 1.7   5    1.6   4  |
|  R21  | 1708  |    .23   .05  | 1.0   0     .9  -1  |
|  R12  | 1568  |    .31   .05  |  .7  -9     .6  -9  |
|  R10  |  731  |    .35   .07  |  .9  -2     .9  -2  |
|  R17  |  113  |    .40   .18  | 1.0   0    1.0   0  |
|  R18  |   49  |    .43   .27  | 1.3   1    1.2   1  |
|  R02  |   13  |   1.02   .53  | 1.2   0    1.1   0  |
|  R16  |  122  |   1.33   .18  | 1.0   0     .9   0  |
|-----------------------------------------------------|
|       |       |          Model |  Infit     Outfit   |
| Rater |       | Measure  S.E. |MnSq ZStd  MnSq ZStd |
|-----------------------------------------------------|
|  Mean | 1235.8|    .00   .13  | 1.1  -1.5  1.0  -2.1|
|  S.D. | 1702.7|    .52   .11  |  .4   4.3   .4   4.2|
|-----------------------------------------------------|
|  RMSE (Model)  .17           Adj S.D.       .49     |
|  Separation   2.86           Reliability   .89      |
|  Fixed chi-square: 462.3     d.f.:          23      |
|                              significance: .00      |
-------------------------------------------------------
```

Discussion

The raters are not all comparable in severity; the analysis indicates that they can be divided into two, almost three, different levels. This can be seen clearly in Figure 5.2, which shows the raters arranged from the most lenient to the most severe, with the standard errors for their measures accounted for. As can be seen, the eight most lenient raters form one difficulty level of their own, as the range of their possible true estimates overlap with each other and do not overlap with that of the other 16 raters. The 16 raters form one group partly because of the large standard error associated with rater R02, the second most severe rater, who provided only 13 ratings. Were R02 to be excluded, the raters would be unambiguously divisible into three levels of severity. On the other hand, if the most severe rater were to be excluded, the raters would then be divisible into only two levels of severity. Raters R02 and R16 are clear outliers in this group. With them, the range of raters' severity spans 2.22 logits. Without them, the range is reduced by more than 40% to 1.32 logits.

Figure 5.3 Ranges of Rater Estimates, Arranged According to Severity



The raters included in this analysis did not all work at the ELI at the same time or rate together with one another. Results to be presented later in Question 5 further show that among raters who actually rated together, the range of their severities is generally lower than 1logit. This difference of severities is less than half the narrowest scale point, meaning that differences in severities are not large enough to affect scores.

In any case, unlike with prompts differing in difficulty, raters' differing in severity is not as much of a cause for concern, unless the range is especially large. The reason being, each composition is read by two raters, and where their ratings differ by more than one scale point, a third rater adjudicates, and the discrepant rating discarded. (The estimates in this analysis included all raters' ratings, including discrepant ratings.) Raters' severities would have to be wildly discrepant—for example, three raters whose severities are each one scale point away from each other—to make agreement on an appropriate score impossible. But as it is, the range of raters' severities here is less than one scale point. Thus, the mechanism of double and triple rating is sufficient in this case to ensure that there is no rater effect as a result of rater severity.

Another category of possible rater error is central tendency. As the name implies, central tendency is the kind of error where raters use just the middle parts of the scale and not its extreme ends. The central tendency error can be investigated by looking at infit and outfit mean square statistics, which have an expected value of one, and a range of zero to positive infinity. Overfit, or having fit values much lower than one, indicates insufficient variation in rater ratings, and the presence of central tendency error. In this analysis, for both infit and outfit, no rater fell below the suggested lower-bound value of 0.4. It would thus appear that there are no errors associated with central tendency.

The fit statistics did show two raters with the opposite problem of underfit, or a lack of consistency in rating. Raters R04 and R13 had fit statistics of 2.6 and 1.6, which are both beyond the suggested upper-bound value of 1.5. One thing that these two raters have in common and which might partly explain their lack of consistency is their combination of inexperience in terms of length of time rating and in terms of compositions rated over a concentrated period of time (cf. Table 4.4). An explanation would be needed, however, for why other raters with similar profiles of inexperience (e.g., R17) show acceptable fit. (The matter of experience will be explored further in the next section.)

In a regime of double marking, ideally, there should be no more than one inconsistent rater. Otherwise, there is the possibility of two inconsistent raters reading the same composition and happening to be inconsistent in the same direction, which would not trigger a third reading, and thus result in an inappropriately higher or lower score for that composition. Where there are several inconsistent raters, the best thing to do would be to make sure that one of the two readings is done by a consistent rater. In that way, any inappropriate rating as a result of rater inconsistency would be discovered and discarded after the third rating is given. In this data, while the two inconsistent raters did overlap with each other, there was no instance where they read the same composition. This would imply that there are no invalid scores in this data as a result of rater inconsistency.

Finally, a third category of possible rater error that can be investigated is restriction of range. (A fourth category, the halo effect, does not apply to and cannot be investigated in assessments that use a single/holistic scale.) Restriction of range is the inability to distinguish test takers into different levels of performance and ability. There are several pieces of evidence that, taken together, suggest that this error is not present in these data. First are the raters' fit statistics previously discussed, which showed that none of the raters overfit. This indicates that ratings have a degree of dispersal across the scale. This is further confirmed by looking at the overall scale (Figure 5.1), which showed that test takers were spread out across the ability range, and that all the scale points were used. Finally, there is the rating scale constructed by the raters' ratings. For the rating scale (and the ratings based on it) to be appropriate, at least two things need to be true. Each scale point must be wide enough so that it represents a distinct ability level, but each scale point must also not be so wide that it covers multiple ability levels. The latter would represent a restriction of range. Linacre (1997) suggests that for the two conditions to be met, the range covered by each scale point should be greater than 1.4 logits but no more than 5.0 logits. Figure 5.1 shows that that is the case with the current rating scale, with the exception of scale point 1. The range of that scale point is 5.33 logits, but given its standard error of 0.36, the scale point's range can actually be less than 5 logits. As it is far from a decision point, and as the MELAB is a test of advanced-level English, dividing scale point 1 into two levels would be of dubious utility. On the whole, it would be safe to say that restriction of range is not a problem with these raters and these ratings.

To summarize, this section addressed the general question regarding the appropriacy and consistency of raters' ratings, and whether or not a rater effect is present in the test. The results of the analysis indicated that differences in rater severity were minor, as were issues of rater consistency, and it also showed that the system of double marking (and third readings whenever required) was sufficient to ensure that these differences did not result in a rater effect. The study, however, still needs to consider two more questions regarding raters before it can conclude that there is no rater effect in this test.

Rater Characteristics

The fifth research question considers the relationship between three rater characteristics and rating quality: **To what extent can it be shown that there is no rater effect as a result of experience, time, and language background?** The first two—rater experience and rater consistency over time—were investigated by dividing the data into three month periods. Following procedures used by Lumley (2006), separate runs of FACETS produced a series of rater severity estimates and fit statistics for each rater. These estimates and statistics were then analyzed for patterns, consistencies, and changes. Because of the nature of the data, a longitudinal study covering four years and all raters was not possible. Instead, two separate periods involving 21 months and 15 months respectively are presented, each involving at least six raters, and each involving two novice raters. The effect of language background on ratings, for its part, was investigated by examining the rater measurement report and the bias/interaction analysis between rater and test-taker language background. As with the bias/interaction analyses in Question 3, attention was paid to the overall chi-square statistics, and to measures and statistics associated with individual bias terms. The results for rater experience and consistency are presented first, followed by the results for rater language background.

Results: Rater Experience and Consistency Over Time

The first period of time studied extended from September 2004 to May 2006, a period of seven quarters. Among the six raters included in the analysis, raters R09 and R23 were beginning raters at the start of the period covered. The raters' severities across the seven quarters of Time Period 1 are presented in Table 5.9, arranged according to severity in the first quarter, while their infit mean square statistics are presented in Table 5.10.

Table 5.9 Raters' Severities Across Quarters, Time Period 1

	Quarter								
	1	**2**	**3**	**4**	**5**	**6**	**7**		
R23*	-1.57	-1.23	-0.60	-0.58	-0.53	-0.29	-0.29		1.28
R01	-0.34	-0.10	-0.32	-0.24	-0.20	-0.09	-0.05		0.29
R12	-0.07	0.28	0.19	0.21	0.20	0.15	0.26		0.35
R21	0.08	-0.21	-0.02	-0.03	0.04	0.15	0.13		0.36
R09*	0.85	0.53	0.35	0.25	0.26	0.11	0.04		0.81
R11	1.05	0.73	0.39	0.40	0.24	-0.04	-0.09		1.14
	2.62	1.96	0.99	0.98	0.79	0.44	0.55		**Range**

Table 5.10 Raters' Infit Across Quarters, Time Period 1

	Quarter								
	1	**2**	**3**	**4**	**5**	**6**	**7**		
R23*	1.3	1.3	1.0	0.9	0.9	0.8	0.8		0.5
R01	0.7	0.7	0.7	0.7	0.8	0.8	0.8		0.1
R12	0.6	0.5	0.5	0.5	0.5	0.6	0.6		0.1
R21	0.9	0.8	0.8	0.8	0.8	0.8	0.8		0.1
R09*	0.6	0.6	0.6	0.6	0.6	0.6	0.6		0.0
R11	0.8	0.7	0.8	0.8	0.7	0.7	0.7		0.1
	0.7	0.8	0.5	0.4	0.4	0.2	0.2		**Range**

Time Period 2 covered the time between November 2006 to January 2008, or a period of five quarters. Seven raters were included in this analysis, and among them, raters R07 and R10 were new raters at the beginning of the period covered. The raters' severities across Time Period 2 are presented in Table 5.11, again arranged according to severity in the first quarter, while their infit mean square statistics are presented in Table 5.12.

Table 5.11 Raters' Severities Across Quarters, Time Period 2

	Quarter					
	1	**2**	**3**	**4**	**5**	
R23	-0.27	-0.51	-0.47	-0.52	-0.58	0.31
R20	-0.08	-0.20	0.07	0.11	0.19	0.39
R09	-0.06	0.07	0.20	0.17	0.08	0.26
R22	0.06	-0.27	-0.19	-0.20	-0.25	0.33
R01	0.09	0.04	0.10	0.14	0.14	0.10
R07*	0.12	0.50	0.03	-0.01	0.06	0.51
R10*	0.14	0.36	0.26	0.30	0.35	0.22
	0.41	1.01	0.73	0.82	0.93	**Range**

Table 5.12 Raters' Infit Across Quarters, Time Period 2

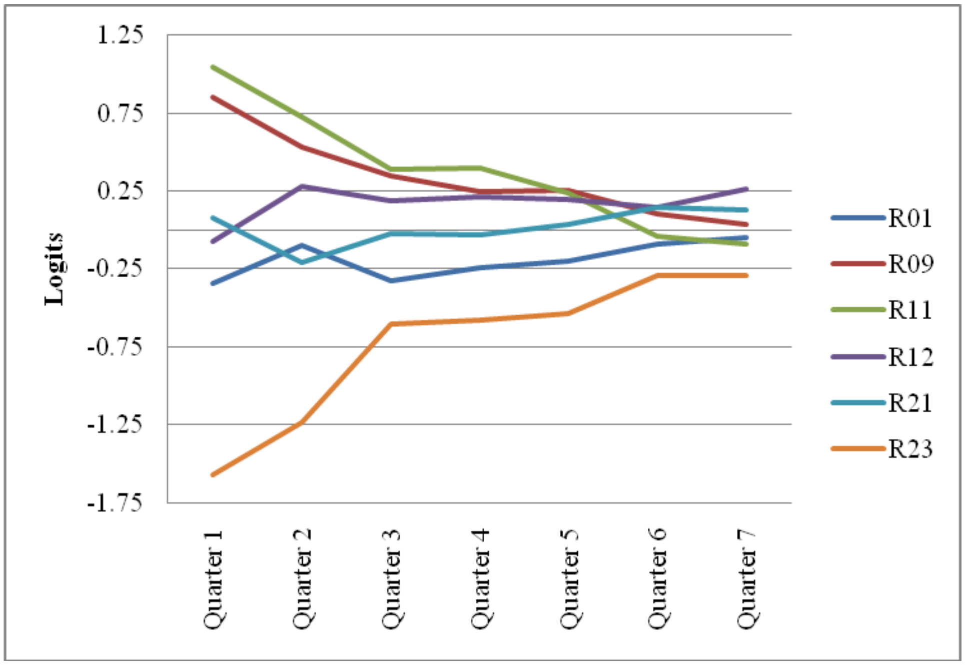
	Quarter					
	1	**2**	**3**	**4**	**5**	
R23	0.5	0.5	0.5	0.5	0.6	0.1
R20	0.5	0.6	0.6	0.6	0.6	0.1
R09	0.5	0.5	0.5	0.6	0.6	0.1
R22	0.4	0.6	0.6	0.6	0.5	0.2
R01	0.7	0.9	0.9	0.9	0.9	0.2
R07*	1.1	0.9	0.8	0.9	0.8	0.3
R10*	1.7	0.8	0.7	0.8	0.7	1.0
	1.3	0.4	0.4	0.4	0.4	**Range**

<u>Discussion: Rater Experience and Consistency Over Time</u>

In many ways, the results presented here are more interesting and more meaningful than those presented in relation to Question 4. First, while the earlier results did involve more raters, the raters did not all overlap with or rate with one another. The results here involve fewer raters—six raters and seven raters, respectively—but involve raters who rated with one another and who did the majority of ratings in the respective time periods covered. It goes without saying that these raters actually needed to agree with one another. As well, noise from raters who rated a few compositions once in a while is excluded. Second, these results also provide a lot more detail, given that there are severity measures and fit estimates for each rater for each three-month period of time. While the overall view gives some information, the focused view can give more. And finally, the detailed view also provides the longitudinal perspective, allowing for the analysis of rater behavior over time.

The first part of Question 5 has to do with rater experience. In Time Period 1 (September 2004 to May 2006), the new raters were R09 and R23. From Table 5.9, it can be seen that at the beginning of Time Period 1, raters R09 and R23's were more severe and more lenient, respectively, relative to the other raters. Rater R11 was also quite severe, and his case deserves further discussion. While R11 had had almost one year of rating experience at the beginning of Time Period 1, he did very few ratings in July and August 2004, the two months immediately prior. In analysis not reported above, R11's severity estimate was 0.18 logits for the first quarter of 2004 and -0.10 for the second quarter of 2004, both of which were very close to the average. From this information, it would appear that the two-month interregnum resulted in R11 rating like an inexperienced rater when he started rating more intensively again in September 2004. This would appear to confirm Lumley & McNamara's (1995) finding—though in their case, among raters of speaking—that training effects tend not to last. Treating rater R11 as a “new” rater along with R09 and R23, a clear trend can be seen in Time Period 1 (Figure 5.3). Over time, the new raters become more and more moderate, until their severity becomes similar to that of more experienced raters. While the range of severities stood at 2.62 logits in Quarter 1, it had been reduced to less than 1 by Quarter 3, and averaged 0.5 by the end of Time Period 1 (cf. narrowest scale point = 2.94 logits).

Figure 5.4 Raters’ Severities Over Time Period 1

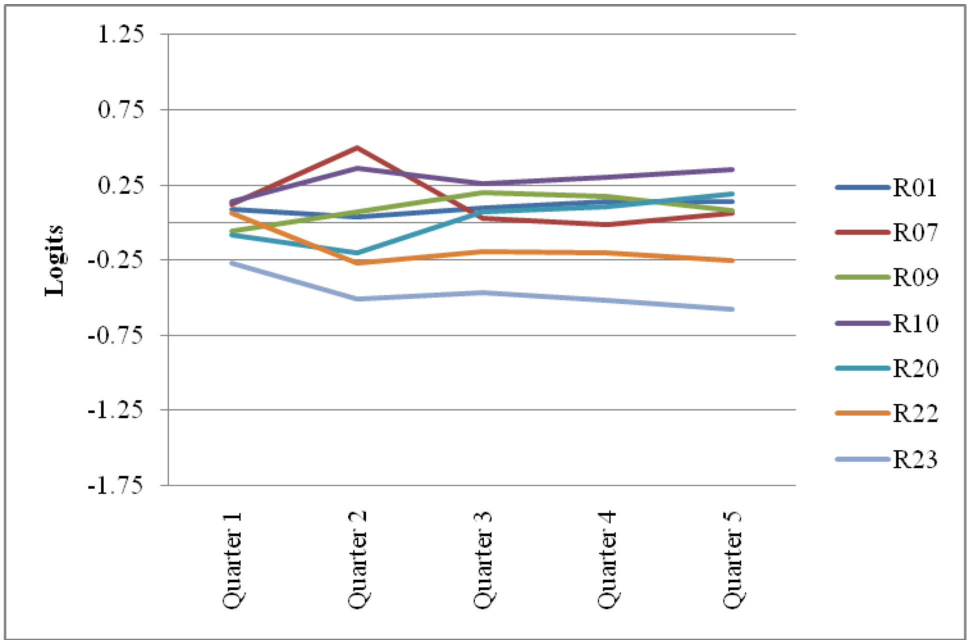


Where consistency is concerned, rater R23 was somewhat more inconsistent at the beginning of Time Period 1, but still within acceptable limits (Table 5.10). Over time, R23 learned to reduce the amount of variation in her ratings until she was at the average for the group. Rater R09, for his part, remained self-consistent throughout the time period, registering an unchanged infit statistic of 0.6 throughout. At the beginning of

Time Period 1, because of R23, the range of raters' infit was 0.7. By the end of the time period, the range had settled at 0.2.

In Time Period 2 (November 2006 to January 2008), the new raters were R07 and R10. The two new raters were the most severe raters in Quarter 1 of this time period. However, their severity estimates were not very far from the average. Both new raters' severity changed somewhat in Quarter 2, after which they appeared to settle on a general level of severity they were each comfortable with (Figure 5.4). Overall, the spread of rater severities was not very large, generally hovering just below 1 logit. As for consistency, raters R07 and R10 both had infit statistics above 1, the latter at 1.7, which is above the limit of acceptability (Table 5.12). It can be seen however that by Quarter 2, both raters had infit values below 1 and stayed below it until the end of the time period.

Figure 5.5 Raters' Severities Over Time Period 2



To summarize the first part of Question 5, a few things can be said about new raters. First, where rater severity is concerned, the data show that new raters' may or may not be very different from more experienced raters. It would appear though that where they are far from the average, they learn to moderate their ratings relatively quickly. As part of their training, new raters are required to keep track of their agreement rates and receive feedback on their rating behavior from other raters. This feedback presumably plays a role in moderating their rating behavior, though how exactly cannot be ascertained by this study. Unlike in most other large-scale assessments, the raters in this test all work in the same location, and while rating particular compositions is done independently, informal and general discussions about the enterprise of rating does happen among the raters. The extent to which this socialization helps moderate raters' rating behavior also cannot be ascertained. Second, where rater consistency is concerned, the data suggests that there is generally more unexpected variation in new raters' ratings. In this data, only new raters had infit statistics greater than one. Three of the four raters had infit statistics of one or lower by their second quarter rating, while the fourth (R23) achieved the same by her third quarter rating. The difference might lie in the volume of reading they did. In their respective second quarters, raters R07, R09, and R10 read over 100 compositions each, with R09 reading 427 for the time period. By contrast, R23 read only 61 compositions in those three months, or less than a composition a day. This information suggests that experience should ultimately be measured not just by time but by rating volume and continuity as well.

One of the raters, R10 (personal communication), suggests a possible explanation for why new raters appear to be more severe and more inconsistent: In rater training, raters are shown training and benchmark compositions at all scale points. Thus, when they begin to rate compositions, they expect to see compositions across the entire scale and rate accordingly. However, in the actual population, there are very few test takers at the extreme ends of the scale. At a certain point in time, new raters realize that this is the case, essentially narrowing their scale, resulting in greater rating consistency. If this explanation is correct, it would suggest the need for raters in training to be given more information about the actual test population.

Evidence from four new raters is probably not basis enough to generalize to new raters in general. On the other hand, the findings here generally accord with the findings in the literature. Weigle (1998) found that inexperienced raters were more severe and less consistent in their ratings compared to experienced raters. Other studies found essentially similar results (Furneaux & Rignall, 2007; Weigle, 1999). Together, the findings of this and other studies provide a clearer picture of the role and effects of experience in rating quality.

The second part of Question 5 has to do with raters' consistency over time. In Lumley's (2006) study, where rater severity is concerned, the four raters were able to maintain their rank order over a period of almost two years. The same is not the case in this study, where raters' rank orders do change (cf. Figures 5.3 & 5.4). There are at least two reasons for these divergent findings. One is that Lumley included only experienced raters in his study, whereas the present study includes inexperienced raters, whose severity tends to show more variability. Another reason is the greater spread of severity among Lumley's raters. The difference between the most severe and the most lenient rater in Lumley's study was approximately 2.25 logits. In the present study, the spread is generally smaller, making it more likely that they change their rank orders.

It will be seen in Table 5.9 that, excluding new raters, the range of individual raters' estimates over the seven quarters of Time Period 1 is no more than 0.36 logits. In Time Period 2 (Table 5.11), the greatest change is 0.39 logits. This is only about a tenth of the range of the average scale point (3.88). Where new raters are concerned, after six months, their severities also become relatively stable. For raters R09 and R23, the range becomes no greater than 0.31 logits. Raters' R07 and R10 show even greater stability in their severities; after six months, the variation is no greater than 0.07 and 0.09, respectively. The raters in this study are also very consistent. Again, apart from the new raters, the range of their infit values is never more than 0.2. As for new raters, they achieve the same stability of consistencies after six months of rating.

The current study is different from most other longitudinal studies of rating quality (e.g., Congdon & McQueen, 2000; Cho, 1999) in that other studies were under experimental conditions. In Congdon and McQueen, for example, rater behavior was shown to change on a day to day basis. However, in their study, raters were asked to read 173 compositions each day, where fatigue likely becomes an extraneous factor. The current study has the advantage of having operational data, and thus provides a better indication of rating quality in real rating situations.

Taken together, it can be concluded that once they gain experience, raters have relatively stable severities and are remarkably consistent in their ratings. There is some indication, however, that continuity is a component of experience, and that the stability and consistency gained with experience can be lost should a rater stop rating for a period of time.

Results: Rater Language Background

The relationship between rater language background and rating quality can partly be investigated through the rater measurement report (Table 5.8). In addition, a relationship between raters and test takers of identical first language backgrounds was investigated through a bias/interaction analysis. Table 5.13 presents bias terms for each rater and his or her native language background, whether or not the bias terms were significant. Three raters—R02, R04, R17—did not rate any compositions written by people who shared their first language, i.e. English. The majority of the bias terms were in the direction of favoring test takers whose first language backgrounds are identical to those of raters. However, only three of the bias terms were significant. Rater R06 favored compositions written by native English speakers by a substantial 0.59 of a scale point. R09 favored Chinese first-language test takers by a minimal 0.04 of a scale point, and R21 showed similarly slight bias (0.11) in favor of Korean first-language test takers.

Table 5.13 Raters and Test-Takers with Identical L1s

```
-----------------------------------------------------
|     Language      |         Obs-Exp|        |Infit |
|     x  Rater      |   n     Average| Z-Score| MnSq |
|---------------------------------------------------|
| English   x R01  |  201      .04  |  -1.15 |  1.1 |
| English   x R03  |    3      .53  |  -1.64 |   .0 |
| Spanish   x R05  |    3      .45  |  -1.53 |  1.0 |
| English   x R06  |    8      .59  |  -3.09 |   .3 |
| English   x R07  |   43     -.01  |   0.10 |  1.1 |
| Spanish   x R08  |    9      .09  |  -0.51 |   .5 |
| Chinese   x R09  | 1119      .04  |  -2.92 |   .7 |
| Filipino  x R09  |  558      .00  |  -0.04 |   .6 |
| English   x R10  |   18      .04  |  -0.35 |   .7 |
| English   x R11  |   79      .05  |  -0.83 |  1.0 |
| English   x R12  |   35      .12  |  -1.32 |  1.1 |
| English   x R13  |    2      .04  |  -0.09 |   .8 |
| English   x R14  |   23      .21  |  -1.83 |   .6 |
| English   x R15  |    3      .27  |  -0.86 |   .3 |
| English   x R16  |    1      .14  |  -0.29 |   .0 |
| English   x R18  |    3     -.02  |   0.06 |  1.1 |
| English   x R19  |    4      .13  |  -0.50 |   .9 |
| English   x R20  |   99     -.06  |   1.09 |  1.4 |
| Korean    x R21  |  104      .11  |  -2.20 |   .8 |
| English   x R22  |  126     -.01  |   0.14 |   .6 |
| English   x R23  |   37      .12  |  -1.40 |   .5 |
| English   x R24  |    2     -.09  |   0.25 |   .0 |
|---------------------------------------------------|
|     Language      |   n     Obs-Exp| Z-Score|Infit |
|     x  Rater      |         Average|        | MnSq |
-----------------------------------------------------
```

Discussion: Rater Language Background

In the field of language testing, it is still an outstanding question whether second language speakers ought to serve as raters in language assessment (Hamp-Lyons & Davies, 2008; Hill, 1996; Reed & Cohen, 2001). Two possible errors are cited. First, it is contended that second language speakers generally rate differently than native speakers do, that they are more severe and more likely to show overfit (Brown, 1995). Second, it has been hypothesized that non-native English speakers could come from places with well-developed varieties of English, which might cause them to overlook or accept features that are unacceptable in a standard dialect.

This study included four raters for whom English was not a first language: raters R05, R08, R09, and R21. Looking at the rater measurement report (Table 5.8), if one were not given information about raters' language background, it would be very difficult to identify which ones were native English speakers and which ones were not. In terms of severity, out of 24 raters, the non-native raters ranked fourth, seventh, twelfth, and eighteenth. With the exception of R08, they were all within one standard deviation of the mean, in the middle of the group. In terms of consistency, among the non-native raters, R09 showed the least amount of variability (0.7). However, two other raters had the same fit statistic, and another rater showed even less variation at 0.6. On the other hand, R08 showed more variation than expected with an infit statistic of 1.4. Thus, the observation that non-native raters are more likely to overfit (Brown, 1995)—which study involved raters of speaking—is not sustained in this study of raters of writing.

Whether raters are more lenient towards test-takers who share their first language background can be seen in Table 5.13. Other than the four non-native raters, data are also shown for the other raters and native English-speaking test takers. The results that involve native English-speaking test takers needs to be interpreted with caution though. In a study using much the same data as this one, it was shown that the ability estimates for these test takers are underestimated (Johnson & Lim, 2009). Thus, the data are more likely to show raters being too lenient to native English-speaking test takers, where in fact that is not the case. Among the 21 raters who rated test takers from their first language background, only three showed significant bias. Rater R06 showed significant and substantial bias for English first-language test takers (0.59)—but as has been said, because these test-takers' abilities are known to be underestimated, the bias measure is therefore an overestimate, and is thus not a valid finding. R09 showed bias of 0.04 or less than one-twentieth of a scale point for Chinese first-language test takers, while R21 showed a bias of 0.11 or approximately one-tenth of a scale point for Korean first-language test takers. The Spanish first-language raters, R05 and R08, did not show significant bias for or against Spanish first language test takers. However, this absence of finding might partly have to do with the small number of compositions written by Spanish speakers that each read.

It must be admitted that the number of non-native raters in this study is relatively small, as is the number of language backgrounds they represent. As well, these raters were all highly proficient in English. Thus, how generalizable these findings are to other non-native raters and to raters of different proficiency levels remains an open question. But in the case of this test and this data, it can be seen that any significant bias shown by non-native raters towards those who share their first language tended to be minimal and not of a magnitude where final scores are affected. Thus, it would appear that there is no rater effect as a result of rater language background in this performance assessment of writing.

Rater Perceptions

Question 6 is concerned with the way that raters' perceptions of prompts and test taker behavior affect their rating behavior: **To what extent can it be shown that raters do not alter their rating behavior depending on perceived differences in prompt difficulty or perceived proficiency-related prompt selection behavior among test takers?** This question was answered first by looking at the responses of the raters to

Forms A and B (Appendices E and F), which provided information about raters' perceptions of relative topic difficulty and raters' perceptions of prompts that lower-level test takers are more likely to choose. Then, a bias/interaction analysis between rater and prompts were conducted, and the results of that analysis—in particular, the differences between observed and expected scores—mapped onto raters' responses to Forms A and B. ANOVAs were conducted to see whether significant differences were present in the scores given to prompts of different perceived levels of difficulty and t-tests were performed to detect significant differences between prompts that lower-level test takers are perceived as more likely and not more likely to choose.

Results

Table 5.14 aggregates the responses of ten raters to Form A, which presented them with 30 prompts and, for each, to respond to the following question: Compared to the average prompt in the pool of MELAB writing prompts, is this prompt easier, about average, or more difficult to get a high score on? The table showed raters' responses to be fairly normally distributed, with 45% of prompts judged to be about average in difficulty, 20.33% and 26.33% for the somewhat easier and somewhat more difficult categories, and 3.67% and 5.67% for the clearly easier and clearly more difficult categories. The overall average for this sample of 30 prompts was 3.08, or just slightly more difficult than the total pool of MELAB prompts, at least according to these raters. For these 30 prompts, there was no correlation between these raters' judgments of prompt difficulty and the actual difficulty (cf. Table 5.1) of the prompts (r = -0.13).

Table 5.14 Raters' Perceptions of Prompt Difficulty (n raters = 10)

Prompt	Raters Who Judged Prompt To Be					Average
	Clearly Easier (=1)	Somewhat Easier	About Average (=3)	Somewhat More Difficult	Clearly More Difficult (=5)	
12	4	2	1	2	1	2.4
44	2	4	2	2	0	2.4
49	1	4	4	1	0	2.5
18	0	4	6	0	0	2.6
22	1	3	5	1	0	2.6
30	1	5	1	3	0	2.6
4	1	3	4	5	0	2.7
46	0	4	5	1	0	2.7
10	0	4	4	2	0	2.8
13	0	4	4	2	0	2.8
16	0	4	3	3	0	2.9
25	0	4	3	3	0	2.9
48	0	3	5	2	0	2.9
54	0	1	9	0	0	2.9
56	0	1	9	0	0	2.9
60	0	2	6	2	0	3.0
32	1	0	6	3	1	3.1
40	0	2	6	1	1	3.1
7	0	1	6	3	0	3.2
37	0	1	7	1	1	3.2
57	0	0	7	3	0	3.3
34	0	0	6	4	0	3.4
38	0	2	2	6	0	3.4
43	0	0	6	4	0	3.4
51	0	0	6	4	0	3.4
5	0	1	3	5	1	3.6
11	0	1	3	4	2	3.7
15	0	1	2	4	3	3.9
14	0	0	3	3	4	4.1
31	0	0	2	5	3	4.1
	Percent of Prompts At Each Level					**Overall Average**
	3.67	20.33	45.33	26.33	5.67	3.08

In Form B, raters were given 24 pairs of prompts and asked to choose, for each pair, which prompt a lower-level test taker was more likely to choose. Seven or more raters selected the same prompt in 12 cases, (with nine raters selecting the same prompt in five of those cases). Those 12 cases are broken down according to different prompt dimensions in Table 5.15, to show the types of prompts that raters think lower-level candidates are more likely to choose. Narrative prompts were selected in four out of five opportunities by this group of raters. This was followed by prompts on personal topics and unconstrained prompts, which were also selected more than 50% of the time.

Table 5.15 Prompts That Raters Think Lower-Level Candidates are More Likely to Choose

Prompt Dimension	Dimension Categories	Number Selected	Number in Sample	Percent Selected
Topic Domain	Business	2	9	22.2%
	Education	1	6	16.6%
	Personal	7	11	63.6%
	Social	2	21	9.5%
Rhetorical Task	Argumentative	0	17	0.0%
	Expository	8	23	34.8%
	Narrative	4	5	80.0%
Prompt Length	1 sentence	1	2	50.0%
	2 sentences	1	12	8.3%
	3 sentences	5	16	31.3%
	4 sentences	5	16	31.3%
	5 sentences	0	2	0.0%
Constraint	Constrained	4	31	12.9%
	Unconstrained	6	11	54.5%
Grammatical Person	1st Person	10	23	43.5%
	3rd Person	2	24	8.3%
Number of Tasks	1	1	7	14.3%
	2	5	17	29.4%
	3	5	19	26.3%
	4	1	3	33.3%

In the bias/interaction analysis between raters and prompts, out of 977 bias terms, (including all 60 prompts and all 24 raters), only a small number showed significant and substantial (i.e. greater than 0.5 scale points) differences between observed and expected scores. Appropriately-measured significant bias terms greater than |0.25| with n-sizes of five or greater are presented in Table 5.16. It can be seen that there are 11 significant and substantial bias terms, three against and eight in favor of test takers. Rater R04 accounted for three of the 11 substantial bias terms. In total, there are almost two times as many bias terms favoring test takers reported here than those that show bias against test takers. One bias term—the case of rater R04 and prompt 24—showed a difference between observed and expected score larger than one scale point.

Table 5.16 Bias/Interaction Analysis: Prompt and Rater

Prompt x Rater		Obs-Exp Average	Z-Score	Infit MnSq	Topic Domain	Rhetor Task	Constr
40	R04	-0.65	3.72	1.5	B	E	U
21	R21	-0.59	3.54	0.8	S	A	C
39	R04	-0.57	3.71	1.2	E	A/E	C
29	R04	-0.49	2.43	0.4	S	A	C
12	R21	-0.46	3.53	1.3	P	E	U
29	R06	-0.44	4.25	0.7	S	A	C
42	R11	-0.42	2.7	0.8	B/S	A	C
24	R15	-0.41	2.14	0.9	S	E	U
27	R08	-0.40	3.64	1.4	S	A	C
31	R07	-0.33	2.77	1.0	S	A	C
52	R12	-0.33	2.2	0.8	E	A	C
12	R23	-0.30	1.96	0.6	P	E	U
22	R21	-0.30	3.56	1.4	P	E	U
28	R07	-0.30	2.14	1.4	S	E	C
28	R08	-0.30	2.81	0.9	S	E	C
58	R09	-0.28	3.54	1.0	E/P	E	C
2	R01	-0.25	2.91	0.7	B	A	C
30	R06	-0.25	2.19	0.8	P	N	U
21	R20	0.26	-4.3	0.8	S	A	C
59	R12	0.26	-2.37	0.8	S	E	C/U
9	R22	0.27	-2.53	0.5	B	E	C/U
18	R11	0.28	-2.17	0.5	S	E	C
30	R12	0.29	-3.36	0.6	P	N	U
38	R09	0.29	-2.41	0.6	S	A	C
47	R06	0.32	-2.77	0.6	S	E	C/U
50	R07	0.33	-2.32	0.7	S	A	C
59	R06	0.33	-3.05	0.9	S	E	C/U
17	R10	0.34	-2.33	0.9	P	E/N	U
31	R10	0.34	-2.42	1.0	S	A	C
43	R11	0.35	-3.49	1.1	S	E	U
51	R15	0.38	-2.13	0.8	S	A	C
59	R23	0.38	-2.36	0.9	S	E	C/U
7	R16	0.43	-2.01	0.4	S	E	C
55	R06	0.44	-3.12	0.9	B	E	C/U
26	R10	0.45	-2.65	0.5	S	A	C
39	R21	0.46	-2.5	0.8	E	A/E	C
55	R08	0.46	-4.51	1.0	B	E	C/U
8	R14	0.47	-3.32	0.7	P	A	C
35	R23	0.48	-2.17	0.2	S	A	C
52	R15	0.48	-2.84	0.9	E	A	C
45	R07	0.52	-3	0.9	B	E	C
2	R10	0.53	-2.74	1.3	B	A	C
27	R10	0.57	-2.61	0.8	S	A	C
6	R07	0.58	-3.14	0.9	B	E	C
26	R23	0.60	-3.02	1.2	S	A	C
22	R17	0.76	-3.66	0.6	P	E	U
25	R06	0.83	-3.73	0.2	P	N	U/C
24	R04	1.14	-5.01	1.0	S	E	U

ANOVAs were conducted for each of the raters who responded to Form A, with perceived prompt difficulty as the categories and the difference between observed and expected scores for the 30 prompts included in Form A as the dependent variable. The results of the ANOVAs are given in Table 5.17. The observed score minus expected score means are in scale point units (i.e. one point is the difference between one scale point and the next). The degrees of freedom differed for the different raters because of differences in the number of categories they checked or because of the number of prompts they actually rated.

Table 5.17 ANOVAs for Prompts Raters Perceive to Be More Difficult

Rater	Observed Score - Expected Score Means					ANOVA			
	Clearly Easier	Somewhat Easier	About Average	Somewhat More Diff.	Clearly More Diff.	df Between	df Within	F	Sig.
R01		.024	-.005	.005	-.030	3	26	.195	.899
R04	-.020	-.330	.210			2	6	.254	.784
R10		-.160	-.046	.027	.040	3	25	.410	.747
R12		.070	.010	.120	.067	3	25	.830	.490
R14	-.155	.090	.118	-.055		3	19	.959	.432
R17		.083	.466	-.040		2	5	.598	.585
R20	.050	.024	-.036	.004	-.110	4	25	.576	.682
R21		.000	-.021	-.050	-.277	3	23	1.356	.281
R22		-.030	.003	.043		2	27	1.196	.318
R23		-.034	-.007	.217	.140	3	24	1.225	.322
Mean	-.042	-.026	.070	.030	-.028				

From the table, a pattern can be seen only in the cases of raters R10, R21 and R22. R10 and R22 become more lenient as prompts become—in their perception—more difficult. R21 showed the opposite tendency, rating prompts she perceives to be more difficult more severely. However, for all ten raters, the results of the ANOVA were not significant. The raters' responses were also recoded into just three categories: easier than average, about average, and more difficult than average. ANOVAs conducted under that coding scheme were also insignificant for all ten raters.

In Form B, raters were asked to indicate which topics lower-level test takers were more likely to choose. T-tests were conducted for each rater, comparing prompts each thought lower-level test-takers were more likely and not more likely to choose. The results show that there were no significant differences between the two classifications of prompts for all ten raters (Table 5.18).

Table 5.18 T-tests for Prompts Raters Think Lower-Level Candidates Prefer

Rater	**Lower-Level Test Takers More Likely to Choose?**			**t-test**	
	Yes	**No**	**Mean Difference**	**t-statistic**	**Sig.**
R01	.002	.007	-.005	-.196	.846
R04	-.123	-.028	-.096	-.251	.806
R10	-.029	-.038	.008	.082	.935
R12	.045	.018	.027	.456	.650
R14	.092	-.053	.145	1.249	.220
R17	.008	.379	-.371	-1.683	.120
R20	.009	-.004	.013	.348	.729
R21	-.058	-.009	-.049	-.676	.503
R22	-.015	.024	-.039	-1.496	.141
R23	.023	.063	-.040	-.546	.588
Mean	-.005	0.036	-.041		

Discussion

Many experts in writing perceive that different rhetorical tasks differ in difficulty. Most think, for example, that narrative writing is easier than argumentative writing. MELAB raters who responded to Form B of this study appear to share the same point of view. When asked to choose prompts that they thought lower-level test takers were more likely to choose, a majority of those prompts they chose involved personal and narrative writing (Table 5.15). Presumably, they think lower-level test takers choose these kinds of prompts because they are easier. However, several studies have found that test takers actually received higher scores on argumentative writing tasks than on narrative writing tasks (Hamp-Lyons & Mathias, 1994; Quellmalz, Capell, & Chou, 1982). Several explanations have been offered for these findings. One is that raters internally adjust their rating behavior depending on how difficult they perceive the writing task to be (Hamp-Lyons & Mathias, 1994; O'Loughlin & Wigglesworth, 2007; Spaan, 1993). That is, they think argumentative tasks are more difficult, and so they rate these more generously, leading to the higher scores observed. Another is that raters perceive lower-level test takers as being more likely to choose narrative tasks, and for this reason, give lower scores to narrative tasks, to match their perception of test-takers' abilities (Wiseman, 2009). Question 6 investigates whether one or the other of these explanations is the reason why scores are the way they are on different types of tasks.

Like other experts, this group of raters appears to think that narrative and personal writing tasks are easier than argumentative tasks. But as was seen in the results of Question 2, scores received by test takers on narrative tasks were not significantly lower than scores they received on argumentative tasks. Neither was there a difference in scores received on first and third person tasks. A more focused way of approaching the question would be to ask individual raters their perceptions of individual prompts' difficulties, and then to see whether they rated prompts of different perceived difficulties differently. This was the approach taken in this study. On average, their bias for the different perceived levels of prompt difficulty was no higher than .07, or less than one-tenth of a scale point. ANOVAs were conducted to see if each rater treated prompts of different perceived levels of ease and difficulty in different ways. The ANOVAs for all ten raters were not significant, indicating that they did not rate prompts of different perceived difficulties differently. Thus, the first explanation, that raters are more generous towards prompts they perceive to be more difficult, does not appear to be correct.

Form B asked the same raters to indicate which prompts lower-level test takers were more likely to choose. If the second possible explanation were correct, prompts that raters perceive lower-level test takers are more likely to choose will be rated lower than other prompts. T-tests for all ten raters were again not significant, indicating that the raters did not systematically give lower scores to prompts that they thought lower-level test takers were more likely to choose. Thus, the second explanation does not appear to be correct, either.

Two other explanations are left, then, for why test takers receive higher ratings (if not necessarily higher scores) on “more difficult” prompts. Hinkel (2002) speculates that simpler prompts tend to result in simpler writing, resulting in lower scores. As this study does not include textual analysis of test-takers’ writing, it cannot argue for or against that speculation. And then there is one final explanation: it could well be that the experts are wrong, and that narrative tasks are actually more difficult than argumentative tasks, and that that is the reason for the difference in ratings. Alternately, it could be that errors that come up more frequently in narrative writing (e.g., tenses) are more salient than those in argumentative writing. This explanation, too, is beyond the ambit of this study. What the study has done is narrow down the possible explanations for the observed differences. To the question of raters adjusting their rating behavior depending on perceived prompt difficulty or on perceived test-taker prompt selection behavior, the results show that they do not change the way they rate for those reasons.

Section Summary

The research questions in this section all dealt with the possibility of a rater effect in writing assessment. That is, that writing test scores might not be valid, reliable, or fair as a result of raters who give inappropriate ratings whether as a result of some general tendency, some background characteristic, or some adjustment of rating behavior based on their perceptions of prompts or test takers. The results of Question 4 indicate that differences in rater severity and issues with rater consistency were minor, and that the differences are more than accounted for by the system of double marking. Question 5 addressed the effects of experience, time, and language background on rating quality. It showed that new raters can at times be more severe and more inconsistent than other raters—but not to such an extent that it could not be addressed by the system of double and triple rating. It also showed that within six months or less, they are indistinguishable from and essentially become experienced raters themselves. The results also showed that the raters were generally stable in their severities and consistency over time. Question 6 provided evidence that raters do not adjust their rating behavior depending on perceived prompt difficulty or perceived test-taker ability. Taken together, the results to the three rater-related questions provide a strong argument that there is no threat to the validity, reliability, or fairness of the test in connection with raters, and that there is no rater effect in this writing performance assessment.

Chapter Summary

In this chapter, I presented the results to the six research questions posed by the study. The questions dealt with the two aspects of writing performance assessments that are usually systematically varied for different test takers: the prompts they respond to, and the raters who rate their responses. The first three questions investigated a range of factors related to prompts that might result in scores that are not valid, including six different dimensions by which prompts may be classified, as well as test-taker gender, language background, and proficiency level. The investigations found that with a few exceptions, perhaps prompts belonging to the social domain, prompts were generally comparable and did not result in prompt-related effects in final scores. Three questions considered the effect of raters on test outcomes, including the possible effects of raters' general rating tendencies, their experience and consistency over time, their language background, and their perceptions of prompts and test takers. The findings showed that the system of double and triple rating more than adequately addressed the minimal differences in rater severity and consistency, and that there was thus no rater effect in the test. Having investigated a large number of factors related to prompts and raters, the study provides a strong argument that providing test takers systematically different treatment in terms of prompts and raters assigned does not create construct-irrelevant variance, and provides evidence in support of writing performance assessments being valid.

CHAPTER 6

CONCLUSION

The previous chapter presented the results of this study which was guided by the question: **How are the validity, reliability, and fairness of a second language writing performance assessment affected by aspects of the examination that are systematically varied for different test takers?** In this chapter, I summarize the findings of the study, situating them in the context of validity investigations in language assessment. I then present the implications of the study for different stakeholders. After that, I put forward some directions for future research to take as a result of the findings, and then conclude this study into the role of prompts and raters in second language writing performance assessment.

Summary of the Study

The question asked by this study has to do with the validity of second language writing performance assessments. In the field of educational measurement in general and in language testing in particular, the consensus is that test validation involves making arguments in support of proposed test interpretations and test uses (Bachman, 2005; Kane, 1992, 2006; Xi, 2007). To that end, an interpretative argument for a composition writing test was sketched out (Figure 2.4) which showed that there were at least five inferences being made between an observation—a writing performance sample—and a decision being made on the basis of that observation. The study concerns the first step of the interpretative argument, the evaluative inference. This step, which has to do with inferring observed scores based on the observations, is sometimes referred to as being about scoring validity, and is considered by some to be the most important step in the interpretative argument (Chapelle, 1999; Shaw & Weir, 2007; Weir, 2005).

The possible threat to score validity identified by the study is the systematic variation typically built into performance writing tests. In particular, different test takers respond to different prompts, and their responses to these different prompts are rated by different raters. Scores can be affected if prompts are not comparable in difficulty or raters are not comparable in severity. As well, there is a problem when any identifiable group's scores are affected by factors that have nothing to do with the construct being measured, as these would indicate the presence of test bias. Prompt and rater effects resulting from these raise issues of validity, reliability, and fairness.

Where prompts are concerned, the results of this study suggest that in second language writing performance assessments such as the MELAB, assigning different prompts to different test takers does not pose a threat to the validity of scores, and that tests are valid, reliable, and fair in that regard. The study found that differences in prompt difficulty did not generally have an effect on scores. Of the many prompt dimensions and test-taker characteristics investigated, only prompts on social topics appeared to be more difficult to a degree that it possibly made a significant difference in scores, and then by only less than 0.15 of a scale point. Excluding a few outlier prompts was suggested to ensure that scores not be unduly affected by prompt variation in every case. The study demonstrated that varying prompts and still having tests that yield valid scores is possible.

Where raters are concerned, the results of this study suggest that raters of second language writing performances such as those in the MELAB can be trained to rate appropriately and consistently, and that under a system of double marking, assigning different raters to different test takers does not pose a threat to the validity of scores, and that tests are valid, reliable, and fair in that regard. The study found that differences in rater severity did not result in inappropriate final scores for test takers. It also found that their ratings were neutral to perceived differences in prompt and test taker characteristics, and that their severity and consistency were generally stable over time. While newer raters exhibited somewhat more variability initially, these were accounted for by double marking, and the process of becoming an experienced rater appears to take a relatively short amount of time of sustained reading.

Taken together, to the question of the effects of systematic variations in test conditions on the validity, reliability, and fairness of second language writing performance assessments, the evidence is strong and the argument is made that varying prompts and raters assigned has no undue effect on score validity. The evaluative inference, the first step in the interpretative argument, can thus be considered to be warranted in the case of the MELAB and of writing assessments like it. Assuming other inferences in the interpretative argument are similarly warranted, then scores interpretations and uses can be considered valid.

Limitations and Generalizability

It cannot be overemphasized that the findings of this study are based on a particular exam with particular features, and employing particular raters working under a particular context. It is thus an open question to what degree the findings apply to other exams and to what extent generalizations can be made. For example, the MELAB employs a small pool of raters, and it is not difficult to argue that it is easier to get a small number of people to share a common understanding of a rating scale than it is to get a large number of people to do the same. Thus, in contexts where a larger pool of raters work independently from different locations, having raters share the same understanding of the rating scale and having raters be relatively comparable in severity might not hold. To cite another example, MELAB raters are all testing professionals working within a testing organization. On the other hand, performance assessments are used in other contexts such as universities, where teachers also serve as raters. The literature indicates that raters' professional backgrounds affect the way they rate (e.g., Brown, 1991; Cumming, et al., 2002; Santos, 1988; Song & Caruso, 1996). Thus, it has to be asked what findings apply and do not apply in other contexts where the participants are different.

In addition, some of the findings were based on limited n sizes. For example, the effects of rater language background on rating behavior was investigated by looking at the rating quality of four raters from just four first-language backgrounds. It remains to be seen whether raters from other first-language backgrounds will perform similarly to those in this study. The four raters in this study were also all highly proficient in English, and the more interesting question might be whether there is a minimum level of proficiency required for those who would rate writing performance. Clearly, as with all other studies, there are limitations to this one. Any generalizations should be made with care, taking into account the specifics of the data and the context from which they are drawn.

Implications

The study's findings has implications for different stakeholders involved in second language writing performance assessment, test users and test providers alike. The general implication for all stakeholders is that where one particular kind of writing performance assessment—the timed, impromptu writing test (Hamp-Lyons, 1991)—is concerned, there can be scoring validity in the sense that scores are not affected by a range of prompt dimensions, rater variables, or test taker characteristics, and that there is no construct-irrelevant variance in that regard.

Test Users

For test users, the implication is that, assuming other steps in the interpretative argument are similarly warranted, test scores can be depended upon to reflect test-takers' writing abilities and can be used to base appropriate decisions on. For test takers in particular, there are implications related to test taking. The knowledge that differences in prompts do not have an appreciable effect on scores should lead them to spend less time worrying about the particular prompt they have been assigned—or in cases where they are allowed a choice of prompt, to spend less time on choosing a prompt—and to spend more of their allotted time actually writing. The combination of less worrying and more writing could potentially lead to washback of better samples of writing, and in turn lead to scores that even more accurately reflect test-takers' writing abilities. Where test takers feel like their test scores are incorrect, they are sometimes allowed to ask that their tests be re-scored, usually for a small fee. This study suggests that paying for a re-score is probably not a wise use of one's money, as scores based on a regime of double marking already discount discrepant ratings and tend to be quite accurate, and a re-score is unlikely to change outcomes.

Test Providers

While test-validation is considered a joint enterprise among all stakeholders (AERA, et al., 1999), in reality the largest part of the responsibility still resides with test providers, and the study's findings has the most implications for them. First, there are implications related to prompts used in writing tests. Prompts in general purpose second language performance tests such as the MELAB are presumed to be accessible to all test takers (Bachman & Palmer, 1996). The study suggests that prompts are indeed robust to differences in test taker gender, language background, and proficiency level. As well, prompts can be allowed to vary according to a number of dimensions without having an effect on their comparability, with the possible exception of topic domain. The results showed that a few prompts were statistical outliers, somewhat more difficult than other prompts were. Because the prompts are secure material, more specific and textual analysis of them is not possible. However, it was seen that a large number of the more difficult prompts were those that dealt with social issues, and a possible significant difference was found between these prompts and prompts on education and business topics. In addition, the eight most difficult prompts were without exception constrained prompts, though the difference between constrained and unconstrained prompts was not significant overall. Thus, it might be appropriate for test providers to exercise special care in the development of prompts belonging to the social domain, perhaps framing the tasks for such prompts so that they are relatively unconstrained. On the other hand, it is also the case that there are social-domain prompts and constrained prompts that are on the easy end of the difficulty scale. What this indicates is that while guidelines exist and can be given for constructing this form of writing stimuli to be generally comparable in difficulty (e.g., Kroll & Reid, 1994), there is probably no fool-proof way of determining how easy or difficult particular prompts will be. Developing prompts is apparently as much art as it is science. The implication here is that trialling of new prompts should always remain a central part of the test development process, so that inappropriately easy or difficult prompts do not get included in live tests. Routine analysis of live prompts also needs to be conducted so that prompts whose statistical performances change, which can indicate that they have been compromised, can be detected and excluded from the pool of prompts.

The suggestion was made in the discussion of the results to exclude some prompts from the pool, in order to ensure comparable difficulty in every case, whether at important decision points or not. However, as prompts on social topics constitute a good number of these prompts, it should be asked what the effects of this action might be. The ability to write on social topics might not be observed often enough, and the ability to extrapolate about it might be hindered. That is to say, to solve the problem of construct-irrelevant variance might in turn create the problem of construct underrepresentation.

Strengthening the evidence for the evaluative inference might weaken the evidence for or alter the inferences regarding explanation and extrapolation. Tests are complex systems; the different parts of an interpretative argument are all related to each other, and actions taken at one level affect and have implications on other levels (Larsen-Freeman & Cameron, 2008). Test providers thus need to keep the whole of the test and the whole of the validity argument in view. This is an especially important consideration for general purpose tests of writing ability that gather a single sample of test-taker writing. Otherwise, they might turn out to be more specific-purpose tests of particular kinds of writing. To account for both construct-irrelevant variance and construct underrepresentation, an option for general purpose tests would be to collect more than one sample of writing, which is the direction a number of exams are taking (e.g., IELTS). Each sample could be more narrowly constrained so that scores can be comparable, while having multiple samples would ensure that the construct is adequately represented. This option, of course, needs to account for the resulting length of the test, and whether that negatively affects performance in any way (e.g., fatigue-related issues), creating different problems yet again.

That differences in the different prompt dimensions mostly had no effect on prompt difficulty has implications for test providers that perhaps also has implications for researchers and theorists of writing and of assessment. The review showed that the enterprise of determining prompt difficulty has been, in a word, difficult (e.g., Norris, et al., 1998). This study has provided additional confirmatory evidence that experts are not necessarily very good at predicting prompt difficulty—as in the example of the relative difficulty of narrative and argumentative tasks (cf. Dobson, Spaan, & Yamashiro, 2003;

Greenberg, 1981; Hamp-Lyons & Mathias, 1994; Mohan & Lo, 1985; Powers & Fowles, 1998). This study also provides evidence that a number of prompt dimensions have no effect on prompt difficulty, thus suggesting that the search for factors need to look in other places. The suggestion from Bachman (2002) to conceive of difficulty not as a property of prompts but as a result of interactions is perhaps the most promising direction to take. Knowing what creates difficulty is central to one task of assessment, that of dividing people into different levels of ability, and a better framework to account for this is essential.

There are also implications for test providers regarding the raters who rate the writing. First, a rater training program similar to the one employed by the MELAB program can produce raters who rate appropriately. As previously described, the multi-stage training program includes guided familiarization from a trainer in the test, the rating scale, and the benchmarks. Calibration ratings lead to monitored live rating, where a new rater keeps track of agreement rates and receives feedback from other raters on their rating behavior. This process goes on until a sufficient volume of ratings at an acceptable level of rating quality is reached, at which point the new rater becomes fully certified. While there clearly are other factors involved, including the nature of the writing task and the rating scale, test programs that adopt similar training programs should find that their raters will similarly be able to rate appropriately.

Second, it is sometimes the case that rating quality is not as good among new raters. While the study does show that any problems related to this are appropriately taken cared of by the system of double marking, and while new raters' learning curves seem to be relatively short, ways of shortening that learning curve further appear to exist.

Appraising raters in training of differences between the actual population of test takers and what the training might suggest can potentially help them rate more moderately and more consistently more quickly when they begin live rating. Increasing the number of compositions they read in the beginning may also hasten the improvement of rating quality.

Third, the results of this and other studies suggest that the effects of training can be lost, and that experienced raters can end up rating like new raters when they stop rating for a period of time. This suggests the need to ensure continuity of rating experience on the one hand, or, if that is not possible, to provide for retraining before these raters rate again.

Finally, while having a few inconsistent raters in the pool of raters does not necessarily have an undue impact on scores in a system employing double marking, it can be a problem when inconsistent raters read the same compositions. It is thus suggested that raters be monitored for consistency, and where inconsistency is observed, to ensure that the second reading be done by a consistent rater, so that inappropriate ratings would be detected and discounted. The ideal, of course, would be to re-train inconsistent raters, or to excuse them from rating if they are unable to show consistency.

Unlike with prompt difficulty and what might account for it, a clear model exists for the rating process, that of Lumley (2006). The model is as thorough as it is well-supported. However, given that it was primarily a cross-sectional study and only of experienced raters, the framework is silent on the matter of experience—how it is gained and what difference it makes in the rating process. There are indications in the literature that new and experienced raters follow fundamentally different processes in rating

(Cumming, et al., 2002; Huot, 1993; Wolfe, et al., 1998), and the results of this study point circumstantially in the same direction. The suggestion was also made that formal and informal feedback (including considering what other raters' opinions of a composition might be) might play a part in moderating and changing rating behavior (cf. Knoch, 2009; O'Sullivan & Rignall, 2007). In certain circumstances, rater socialization might also be a factor affecting rating behavior. The implication then is that there is a need not just for a model of the rating process, but also one for rater development, accounting for these longitudinal and social aspects, and how these affect or change the rating process. This is one more thing for test theorists and researchers to pursue.

Directions for Future Research

While this study helped to clarify the role of prompts and raters in second language writing performance assessment, it also raises questions for future research to pursue.

Regarding writing prompts, the study established that variations in prompts writing did not have an undue effect on test-taker scores. For the providers and users of this test, that is clearly a desirable outcome. For theorists and researchers, that might not necessarily be the case. As previously mentioned, the finding of no differences as a result of a range of prompt dimensions and test-taker characteristics means that the search for an explanation for task difficulty continues. One possible reason why differences were not detected is that the variation was actually rather limited, in that the prompts all required the same, one genre of writing. Doubts have been raised regarding the single sample writing test (e.g., Purves, 1992), and a number of language proficiency tests have moved toward multiple samples of writing. This absence of other genres also prevented the study from making a determination regarding construct underrepresentation. Studies are certainly in order that include multiple genres of writing. These studies would enhance the field's understanding and knowledge regarding task difficulty and construct definition.

Regarding raters, the study showed that trained raters generally rate appropriately. That is to say, there appears to be good agreement among the raters. However, the study was not in a position to investigate what it was exactly the raters were agreeing about. More studies in support of the explanatory inference (cf. Figure 2.4) are definitely in order. The study also raised a few interesting questions regarding who should be raters and how raters are socialized and become experienced.

First is the question of who should serve as raters. The study showed that even among new raters there were variations in rating quality, suggesting that some take more naturally to the task of rating than others. What then are the qualities and characteristics of persons who are suited to being raters? The study also showed that not being a native speaker need not be a bar to people serving as raters in that language. In the study, however, the non-native English-speaker raters all had a high level of proficiency in English, which invites the question whether there is a minimum level of proficiency required below which a person is unable to rate writing appropriately or consistently. As well, the non-native raters came from a small number of language backgrounds. Would raters from other language backgrounds perform similarly to those in the study? Thus, there remain important questions regarding who can or should serve as raters, the answers to which have implications for rater recruitment, training, and deployment.

Second is the question of rater experience and socialization. The study showed that new raters took some time, if relatively short, to rate in the same way as experienced raters. The question then is whether raters need to remain in training longer, or if improved rating quality necessarily requires and is the result of actual rating experience. One limitation of the study is that its data only allowed it to track rater development in three-month periods. Other studies that use a smaller unit of time can give more fine-grained information regarding rater development and rating stability over time. The results of the study also indicate that defining experience rating is not an uncomplicated matter. Apart from the length of time one has been rating, experience potentially also consists of quantity of ratings and continuity of experience, both of which apparently have an effect on rating quality. Future research can certainly look into how these different factors of time, quantity, and continuity—as well as individual differences—relate and combine with one another to yield rating quality reflective of "experienced" raters. This can perhaps be done with reference to research in other areas on the nature of expertise (e.g., Ericsson, Charness, Feltovich, & Hoffman, 2006).

Finally, there is the question of rater socialization. The MELAB is distinct in that it is a large-scale testing program that is also at the same time relatively small, allowing a situation where ratings are done by raters in one location. In this kind of setup, there are opportunities for raters to interact and discuss with each other the rating task, adding an extra layer of complexity to rater behavior and ratings. While this setup is not necessarily common in large-scale testing, it is relatively common in other settings—such as schools and universities—where writing performance assessments are also used, making this question one that is worth pursuing. Future research can and should look into the effect of rater socialization on rater behavior and ratings.

Chapter Summary

Writing performance assessments were developed because it was thought that they better reflected writing ability and would result in positive washback for language learners. But as with all assessments, their validity needed to be established. The process of test validation involves making arguments about the interpretations and inferences being made. One part of that process is showing that there is score validity, that scores reflect the ability being measured, and is not affected by factors extraneous to the ability being measured. In writing performance assessments, practical constraints require test takers to respond to different prompts and for their responses to be read by different raters. This systematic variation in treatment that test takers receive pose a threat to the validity, reliability, and fairness of these tests.

This study investigated the effect of these variations on scores in one large-scale assessment. Overall, assigning different prompts and different raters to different test takers did not appear to unduly affect the scores that test takers received. This finding was robust to a range of prompt dimensions, rater variables, and test taker characteristics. It would thus appear that there is score validity in the test and can be for other second language writing performance assessments like it.

The study addressed only one part of the interpretative argument for the validity of a composition writing test. And of the two general threats to construct validity—construct underrepresentation and construct-irrelevant variance—it only addressed the latter. No study can finally establish validity once and for all time, because a definition of validity as making arguments about test interpretations and test uses implies that validity is always a matter of degree and always provisional. So, while this study helped illumine one aspect of the validity, reliability, and fairness of second language writing performance assessment, the work of validation continues.

APPENDICES

APPENDIX A
MELAB WRITING INSTRUCTIONS

MICHIGAN ENGLISH LANGUAGE ASSESSMENT BATTERY

PART 1: COMPOSITION

NAME (PRINT) ______________________________________ Date _______________
(family/surname) (given/first name)

SIGNATURE ____

INSTRUCTIONS:

1. You will have 30 minutes to write on one of the two topics printed below. If you do not write on one of these topics, your paper will not be scored. If you do not understand the topics, ask the examiner to explain or to translate them.

2. You may make an outline if you wish, but your outline will not count toward your score.

3. Write about 1 to 2 pages. Your composition will be marked down if it is extremely short. Write on both sides of the paper. Ask the examiner for more paper if you need it.

4. You will not be graded on the appearance of your paper, but your handwriting must be readable. You may change or correct your writing, but you should not copy the whole composition over.

5. Your essay will be judged on clarity and overall effectiveness, as well as on
   - topic development
   - organization
   - range, accuracy, and appropriateness of grammar and vocabulary

APPENDIX B
MELAB COMPOSITION RATING SCALE

97
Topic is richly and fully developed. Flexible use of a wide range of syntactic (sentence level) structures, accurate morphological (word forms) control. Organization is appropriate and effective, and there is excellent control of connection. There is a wide range of appropriately used vocabulary. Spelling and punctuation appear error free.

93
Topic is fully and complexly developed. Flexible use of a wide range of syntactic structures. Morphological control is nearly always accurate. Organization is well controlled and appropriate to the material, and the writing is well connected. Vocabulary is broad and appropriately used. Spelling and punctuation errors are not distracting.

87
Topic is well developed, with acknowledgement of its complexity. Varied syntactic structures are used with some flexibility, and there is good morphological control. Organization is controlled and generally appropriate to the material, and there are few problems with connection. Vocabulary is broad and usually used appropriately. Spelling and punctuation errors are not distracting.

83
Topic is generally clearly and completely developed, with at least some acknowledgement of its complexity. Both simple and complex syntactic structures are generally adequately used; there is adequate morphological control. Organization is controlled and shows some appropriacy to the material, and connection is usually adequate. Vocabulary use shows some flexibility, and is usually appropriate. Spelling and punctuation errors are sometimes distracting.

77
Topic is developed clearly but not completely and without acknowledging its complexity. Both simple and complex syntactic structures are present; in some "77" essays these are cautiously and accurately used while in others there is more fluency and less accuracy. Morphological control is inconsistent. Organization is generally controlled, while connection is sometimes absent or unsuccessful. Vocabulary is adequate, but may sometimes be inappropriately used. Spelling and punctuation errors are sometimes distracting.

73
Topic development is present, although limited by incompleteness, lack of clarity, or lack of focus. The topic may be treated as though it has only one dimension, or only one point of view is possible. In some "73" essays both simple and complex syntactic structures are

present, but with many errors; others have accurate syntax but are very restricted in the range of language attempted. Morphological control is inconsistent. Organization is partially controlled, while connection is often absent or unsuccessful. Vocabulary is sometimes inadequate, and sometimes inappropriately used. Spelling and punctuation errors are sometimes distracting.

67
Topic development is present but restricted, and often incomplete or unclear. Simple syntactic structures dominate, with many errors; complex syntactic structures, if present, are not controlled. Lacks morphological control. Organization, when apparent, is poorly controlled, and little or no connection is apparent. Narrow and simple vocabulary usually approximates meaning but is often inappropriately used. Spelling and punctuation errors are often distracting.

63
Contains little sign of topic development. Simple syntactic structures are present, but with many errors; lacks morphological control. There is little or no organization, and no connection apparent. Narrow and simple vocabulary inhibits communication, and spelling and punctuation errors often cause serious interference.

57
Often extremely short; contains only fragmentary communication about the topic. There is little syntactic or morphological control, and no organization or connection are apparent. Vocabulary is highly restricted and inaccurately used. Spelling is often indecipherable and punctuation is missing or appears random.

53
Extremely short, usually about 40 words or less; communicates nothing, and is often copied directly from the prompt. There is little sign of syntactic or morphological control, and no apparent organization or connection. Vocabulary is extremely restricted and repetitively used. Spelling is often indecipherable and punctuation is missing or appears random.

N.O.T. (Not On Topic)
Indicates a composition written on a topic completely different from any of those assigned; it does not indicate that a writer has merely digressed from or misinterpreted a topic. N.O.T. compositions often appear prepared and memorized. They are not assigned scores or codes.

APPENDIX C
PROMPT CODES

Legend:

Topic Domain	Rhetorical Task	Task Constraint	Grammatical Person
B: Business E: Education P: Personal S: Social	A: Argumentative E: Expository N: Narrative	C: Constrained U: Unconstrained	1: 1st person 3: 3rd person

Prompt	**Topic Domain**	**Rhetorical Task**	**Task Constraint**	**Grammatical Person**	**Number of Tasks**
1	E/P	A	C	3/1	2
2	B	A	C	3	2
3	E	A/E	C	3	3
4	S	E	U	1	3
5	E	A	C	1	3
6	B	E	C	1	3
7	S	E	C	1	3
8	P	A	C	1	2
9	B	E	C/U	3	2
10	P	N	U	1	1
11	B	E	C	3	2
12	P	E	U	1	2
13	P	E	C	1	3
14	S	A	C	3	2
15	S	E	C	3	2
16	P	N	U	1	1
17	P	E/N	U	1/3	1
18	S	E	C	1	2
19	S	E	C/U	3	1
20	P	E	C	1	2
21	S	A	C	3	2
22	P	E	U	1	3
23	S	E	C	3	4
24	S	E	U	1	3/2
25	P	N	U/C	1	3
26	S	A	C	3	3
27	S	A	C	3/1	3

Prompt	**Topic Domain**	**Rhetorical Task**	**Task Constraint**	**Grammatical Person**	**Number of Tasks**
28	S	E	C	3	4
29	S	A	C	3	3
30	P	N	U	1	3/2
31	S	A	C	3	2
32	S	E	C	3	3
33	S	A	C	3	3
34	S	A	C	1	2
35	S	A	C	3	2
36	S	E	C	3	2
37	B	E	U	3	2
38	S	A	C	3	2
39	E	A/E	C	3	3
40	B	E	U	3	3
41	S	E	C/U	3	4
42	B/S	A	C	3	3
43	S	E	U	3	1
44	P	E	U	1	4
45	B	E	C	3	1
46	P	N	U/C	1	2/5
47	S	E	C/U	3	1
48	B	E	C	1	2
49	B	E	C	3	2
50	S	A	C	1	2
51	S	A	C	1	3
52	E	A	C	1	4
53	S	E	C	3	3
54	E	A	C	3	3
55	B	E	C/U	1	4
56	E	A	C	3	3
57	S	A	C	1	2
58	E/P	E	C	1	2
59	S	E	C/U	3	1
60	S	A	C	3	3

APPENDIX D
INSTRUCTIONS TO RESPONDENTS, FORMS A AND B

Instructions:

1. Please respond to the two instruments in the order presented.

2. Work relatively quickly through the instruments, i.e. there is no need to second guess yourself.

3. Do take a break at any point you feel the need. However, as much as possible, try to work on the instruments under the same general conditions (e.g. mood, place, etc.)

4. If for any reason you don’t feel confident about the quality of your responses to any part, complete the instruments anyway, and then relay to me your concerns.

5. While you are being asked for your name, it will only be used for data matching purposes, and you will remain anonymous in any product of this research.

Thank you again for agreeing to participate in this study.

APPENDIX E
FORM A

Name: ______________________________________

Instructions: Place yourself in the same frame of mind as when you are reading and rating MELAB compositions. For each prompt below, check the box corresponding to your answer to the following question:

COMPARED TO THE AVERAGE PROMPT IN THE POOL OF MELAB WRITING PROMPTS, IS THIS PROMPT EASIER, ABOUT AVERAGE, OR MORE DIFFICULT TO GET A HIGH SCORE ON?

	Clearly easier	**Somewhat easier**	**About average**	**Somewhat more difficult**	**Clearly more difficult**
<Text of Topic 32 appeared here>					
<Text of Topic 34appeared here>					
<Text of Topic 7 appeared here>					
. . .					

APPENDIX F
FORM B

Name: ____________________________________

Instructions: Place yourself in the same frame of mind as when you are reading and rating MELAB compositions. For each pair of writing prompts, the questions you are responding to are:

ARE LOWER-ABILITY LEVEL CANDIDATES MORE LIKELY TO CHOOSE ONE PROMPT OR THE OTHER? IF YES, WHICH ONE?

	No	Yes	→	Which?
<Text of Topic 1 appeared here>				
<Text of Topic 2 appeared here>				
<Text of Topic 3 appeared here>				
<Text of Topic 4 appeared here>				
…				
…				

APPENDIX G
FACETS COMMAND FILE

```
Title = MainRun

Data File = MainRun.dat          ;
Output File = MainRunOutput.txt  ;
Score File = MainRunscore,Tab    ;

Facets = 7              ;
Positive = 1            ;
Noncenter = 1           ;
Iterations = 500        ;
Convergence = 0.5, .01  ;
Unexpected = 3.0        ;
Xtreme = 0.3, 0.5       ;
Zscore = 0,0            ;
Inter-rater = 7         ;

Arrange = 1A,2A,3A,3M,4A,5A,6A,6M,7A,7M,N ;
Vertical = 1*,3A,6A,7A  ;
Lefthand = yes          ;

Model =
?,?B,?,?,?,?B,?,R9      ; prompt x gender
?,?,1-59B,?,?,?B,?,R9   ; prompt x language background
?,?,?,,?B,?B,?,R9       ; prompt x proficiency level
?,?,1-59B,?,?,?,?B,R9   ; rater x language background
?,?,?,?,?,?B,?B,R9      ; rater x prompt
?,?,?,1-5B,?,?,?B,R9    ; rater x perception
?,?,?,?,?B,?,?B,R9      ; rater x proficiency
*
Labels =
1, Examinee             ;
18000-99999
*
2, Gender               ;
1=Male
2=Female
*
3, L1                   ;
1=Alba
2=Amha
3=Arab
4=Arme
5=Beng
6=Bosn
7=Bulg
```

```
8=Camb
9=Chin
10=Dari
11=Engl
12=Fars
13=Fili
14=Fren
15=Germ
16=Gree
17=Guja
18=Hebr
19=Hind
20=Hung
21=Ibo
22=Indo
23=Ital
24=Japa
25=Kaza
26=Kore
27=Mace
28=Maly
29=Mlym
30=Mart
31=Nepa
32=Orom
37=Pash
38=Poli
39=Port
40=Punj
41=Roma
42=Russ
43=Serb
44=SrCr
45=Sinh
46=Slov
47=Soma
48=Span
49=Swah
50=Tami
51=Telu
52=Thai
53=Tibe
54=Tigr
55=Turk
56=Ukra
57=Urdu
58=Viet
59=Yoru
```

```
63=OAfr
64=OAsi
65=OEur
66=OInd
*
4, Perception, A        ;
1-5=,0
*
5, Level, A             ;
4-9=,0
*
6, Prompt               ;
1-60
*
7, Rater                ;
1-24
```

REFERENCES

Alderson, J. C. (1991). Bands and scores. In J. C. Alderson & B. North (Eds.), *Language testing in the 1990s* (pp. 71-86). London: Macmillan.

Allen, N. L., Holland, P. W., & Thayer, D. T. (2005). Measuring the benefits of examinee-selected questions. *Journal of Educational Measurement*, *42*(1), pp. 27-51.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1966). *Standards for educational and psychological tests and manuals.* Washington, DC: American Psychological Association.

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.

American Psychological Association. (1954). *Technical recommendations for psychological tests and diagnostic techniques*. Washington, DC: Author.

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

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

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

Bachman, L. F. (2007). What is the construct? The dialectic of abilities and contexts in defining constructs in language assessment. In J. Fox, M. Wesche, D. Bayliss, L. Cheng, C. E. Turner, & C. Doe (Eds.), *Language testing reconsidered* (pp. 41-71). Ottawa, Canada: University of Ottawa Press.

Bachman, L. F., Lynch, B. K., & Mason, M. (1995). Investigating variability in tasks and rater judgements in a performance test of foreign language speaking. *Language Testing*, *12*(3), pp. 238-257.

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

Barkaoui, K. (2007a). Participants, texts, and processes in ESL/EFL essay tests: A narrative review of the literature. *Canadian Modern Language Review*, *64*(1), pp. 99-134.

Barkaoui, K. (2007b, April). *Raters' perceptions of the effects of thinking aloud on their ESL essay rating performance: A qualitative study*. Paper presented at the annual meeting of the American Association for Applied Linguistics, Costa Mesa, CA.

Barkaoui, K. (2009, March). *The role of scoring method and rater experience in ESL essay rating: A qualitative study*. Paper presented at the annual meeting of the American Association for Applied Linguistics, Denver, CO.

Barnwell, D. (1989). Naïve native speakers and judgments of oral proficiency in Spanish. *Language Testing*, *6*(2), pp. 152-163.

Barritt, L., Stock, P. L., & Clark, F. (1986). Researching practice: Evaluating assessment essays. *College Composition and Communication*, *37*(3), pp. 315-327.

Bell, J. F. (1997). Question choice in English literature examinations. *Oxford Review of Education*, *23*(4), pp. 447-458.

Breland, H., Bridgeman, B., & Fowles, M. (1999). *Writing assessment in admission to higher education: Review and framework*. College Board Report, 99-03. Princeton, NJ: Educational Testing Service.

Breland, H., Lee, Y. W., Najarian, M., & Muraki, E. (2004). *An analysis of TOEFL CBT writing prompt difficulty and comparability for different gender groups*. TOEFL Research Reports, RR-04-05. Princeton, NJ: Educational Testing Service.

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

Brennan, R. L. (2006). Perspectives on the evolution and future of educational measurement. In R. L. Brennan (Ed.), *Educational measurement*, (4th ed., pp. 1-16). Westport, CT: American Council on Education.

Bridgeman, B., Morgan, R., & Wang, M. M. (1997). Choice among essay topics: Impact on performance and validity. *Journal of Educational Measurement*, *34*(3), pp. 273-286.

Broer, M., Lee, Y. W., Rizavi, S., & Powers, D. (2005). *Ensuring the fairness of GRE writing prompts: Assessing differential difficulty*. ETS Research Report, RR 05-11. Princeton, NJ: Educational Testing Service.

Brossell, G. (1983). Rhetorical specification in essay examination topics. *College English*, *45*(2), pp. 165-173.

Brossell, G., & Ash, B. H. (1984). An experiment with the wording of essay topics. *College Composition and Communication*, *35*(4), pp. 423-425.

Brown, A. (1995). The effect of rater variables in the development of an occupation-specific language performance test. *Language Testing*, *12*(1), pp. 1-15.

Brown, A. (2003). Interviewer variation and the co-construction of speaking proficiency. *Language Testing*, *20*(1), pp. 1-25.

Brown, H. D. (1994). *Teaching by principles: An interactive approach to language pedagogy*. Englewood Cliffs, NJ: Prentice Hall Regents.

Brown, J. D. (1991). Do English and ESL faculties rate writing samples differently? *TESOL Quarterly*, *25*(4), pp. 587-603.

Brown, J. D., Hilgers, T., & Marsella, J. (1991). Essay prompts and topics: Minimizing the effect of mean differences. *Written Communication*, *8*(4), pp. 533-556.

Brown, J. D., Hudson, T. D., Norris, J. M., & Bonk, W. (2002). *An investigation of second language task-based performance assessments.* Technical report no. 24. Hawaii: University of Hawaii Press.

Canale, M. (1983). On some dimensions of language proficiency. In J. W. Oller (Ed.), *Issues in language testing research*. Rowley, MA: Newbury House.

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

Carlson, S. B. (1988). Cultural differences in writing and reasoning skills. In A. C. Purves (Ed.), *Writing across languages and cultures: Issues in contrastive rhetoric* (pp. 227-260). Newbury Park, CA: Sage.

Carlson, J. G., Bridgeman, B., Camp, R., & Waanders, J. (1985). *Relationship of admission test scores to writing performance of native and nonnative speakers of English*. TOEFL Research Report #19. Princeton, NJ: Educational Testing Service.

Chalhoub-Deville, M. (2003). Fundamentals of ESL admissions tests: MELAB, IELTS, and TOEFL. In D. Douglas (Ed.), *English language testing in US colleges and universities*, (2nd ed., pp. 11-35). Washington DC: NAFSA.

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

Chapelle, C. A. (1999). Validity in language assessment. *Annual Review of Applied Linguistics*, *19*, pp. 254-272.

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

Charney, D. (1984). The validity of using holistic scoring to evaluate writing. *Research in the Teaching of English*, *18*, pp. 65-81.

Chiste, K. B., & O'Shea, J. (1988). Patterns of question selection and writing performance of ESL students. *TESOL Quarterly*, *22*, pp. 681-684.

Chiswick, B. R., & Miller, P. W. (2004, August). *Linguistic distance: A quantitative measure of the distance between English and other languages* (Discussion Paper No. 1246). Bonn, Germany: Institute for the Study of Labor.

Cho, D. W. (1999). A study on ESL writing assessment: Intra-rater reliability of ESL compositions. *Melbourne Papers in Language Testing*, *8*(1), pp. 1-24.

Clauser, B. E., & Mazor, K. M. (1998). Using statistical procedures to identify differentially functioning test items. *Educational Measurement: Issues and Practice*, *17*(1), pp. 31-44.

Congdon, P. J. & McQueen, J. (2000). The stability of rater severity in large-scale assessment programs. *Journal of Educational Measurement*, *37*, pp. 163-178.

Connor, U., & Carrell, P. L. (1993). The interpretation of tasks by writers and readers in holistically rated direct assessment of writing. In J. G. Carson & I. Leki (Eds.), *Reading in the composition classroom: Second language perspectives* (pp. 141-160). Boston, MA: Heinle and Heinle.

Connor-Linton, J. (1995a). Looking behind the curtain: What do L2 composition ratings really mean? *TESOL Quarterly*, *29*, pp. 762-765.

Connor-Linton, J. (1995b). Crosscultural comparison of writing standards: American ESL and Japanese EFL. *World Englishes*, *14*(1), pp. 99-115.

Cronbach, L. J. (1988). Five perspectives on the validity argument. In H. Wainer & H. I. Braun (Eds.), *Test validity* (pp. 3-17). Hillsdale, NJ: Lawrence Erlbaum.

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin*, *55*(4), pp. 281-302.

Crowhurst, M. (1980). Syntactic complexity and teachers' ratings of narrations and arguments. *Research in the Teaching of English*, *14*(3), pp. 223-231.

Cumming, A. (1990). Expertise in evaluating second language compositions. *Language Testing*, *7*(1), pp. 31-51.

Cumming, A., Kantor, R., & Powers, D. E. (2002). Decision making while rating ESL/EFL writing tasks: A descriptive framework. *The Modern Language Journal*, *86*(1), pp. 67-96.

Cumming, A., Kantor, R., Powers, D., Santos, T., & Taylor, C. (2000). *TOEFL 2000 writing framework: A working paper*. TOEFL Monograph Series, MS-18. Princeton, NJ: Educational Testing Service.

Cureton, E. E. (1951). Validity. In E. F. Lindquist (Ed.), *Educational measurement* (pp. 621-694). Washington, DC: American Council on Education.

Davies, A. (2004). The native speaker in applied linguistics. In A. Davies & C. Elder (Eds.), *The handbook of applied linguistics* (pp. 431-450). Malden, MA: Blackwell.

DeRemer, M. L. (1998). Writing assessment: Raters' elaboration of the rating task. *Assessing Writing*, *5*(1), pp. 7-29.

Diederich, P. B. (1974). *Measuring growth in English*. Urbana, IL: NCTE.

Diedrich, P. B., French, J. W., & Carlton, S. T. (1961). Factors in judgments of writing ability. *Research Bulletin 61*(15). Princeton, NJ: Educational Testing Service.

Dobson, B. K., Spaan, M. C., & Yamashiro, A. D. (2003, July). What's so hard about that? Investigating item/task difficulty across two examinations. Poster presented at the Language Testing Research Colloquium, Reading, United Kingdom.

Dunbar, S. B., Koretz, D. M., & Hoover, H. D. (1991). Quality control in the development and use of performance assessments. *Applied Measurement in Education*, *4*(4), pp. 289-303.

Englehard, G. (1992). The measurement of writing ability with a many-faceted Rasch model. *Applied Measurement in Education*, *5*(3), pp. 171-191.

Englehard, G. (1994). Examining rater errors in the assessment of written compositions with a many-faceted Rasch model. *Journal of Educational Measurement*, *31*(2), pp. 93-112.

English Language Institute, University of Michigan. (2005). *Michigan English language assessment battery: Technical manual 2003*. Ann Arbor, MI: English Language Institute, University of Michigan.

Erdosy, M. U. (2003). *Exploring variability in judging writing ability in a second language: A study of four experienced raters of ESL compositions.* TOEFL Research Reports, RR-03-17. Princeton, NJ: Educational Testing Service.

Ericsson, K. A., Charness, N., Feltovich, P., & Hoffman, R.R. (Eds.) (2006) *Cambridge handbook on expertise and expert performance.* Cambridge, UK: Cambridge University Press.

Ericsson, K. A., & Simon, H. A. (1993). *Protocol analysis: Verbal reports as data*. 2nd edition. Cambridge, MA: MIT Press.

Evans, T. D. (1988). A consideration of the meaning of the word "discuss' in examination questions. In P. C. Robinson (Ed.), *Academic writing: Process and product* (pp. 47-52). Oxford: Modern English Publications.

Fayer, J. M., & Krasinski, E. (1987). Native and nonnative judgements of intelligibility and irritation. *Language Learning*, *37*(3), pp. 313-326.

Fisher, A. G. (1993). The assessment of IADL motor skills: An application of many-faceted Rasch analysis. *American Journal of Occupational Therapy*, *47*(3), pp. 319-329.

Fitzpatrick, A. R., Ercikan, K., Yen, W. M., & Ferrara, S. (1998). The consistency between raters scoring in different test years. *Applied Measurement in Education*, *11*(2), pp. 195-208.

Fitzpatrick, A. R., & Yen, W. M. (1995). The psychometric characteristics of choice items. *Journal of Educational Measurement*, *32*(3), pp. 243-259.

Freedman, S. W. (1983). Student characteristics and essay test writing performance. *Research in the Teaching of English*, *17*(4), pp. 313-325.

Freedman, S. W. (1984). The registers of student and professional expository writing. Influences on teachers' responses. In R. Beach & S. Bridwell (Eds.), *New directions in composition research* (pp. 334-347). New York: Guilford Press.

Freedman, S. W., & Calfee, R. C. (1983). Holistic assessment of writing: Experimental design and cognitive theory. In P. Mosenthal, L. Tamor, & S. A. Walmsley (Eds.), *Research on writing: Principles and methods* (pp. 75-98). New York: Longman.

Fremer, J., Jackson, R., & McPeek, M. (1968). *Review of the psychometric characteristics of the Advanced Placement Tests in Chemistry, American History, and French*. Princeton, NJ: Educational Testing Service.

Fulcher, G., & Davidson, F. (2009). Test architecture, test retrofit. *Language Testing*, *26*(1), pp. 123-144.

Fulcher, G., & Reiter, R. M. (2003). Task difficulty in speaking tests. *Language Testing*, *20*(3), pp. 321-344.

Furneaux, C., & Rignall, M. (2007). The effect of standardization-training on rater judgements for the IELTS Writing Module. In L. Taylor & P. Falvey (Eds.), *IELTS collected papers: Research in speaking and writing assessment* (pp. 422-445). Cambridge: Cambridge University Press.

Gabrielson, S., Gordon, B., & Englehard, G. (1995). The effects of task choice on the quality of writing obtained in a statewide assessment. *Applied Measurement in Education*, *8*(4), pp. 273-290.

Gass, S., Winke, P., & Reed, D. (2007, June). The effect of rater background on the evaluation of speech samples. Paper presented at the Language Testing Research Colloquium, Barcelona, Spain.

Gee, T. W. (1985). Of ships and sealing wax: Drafting and revising processes in grade twelve students' examination writing. *English Quarterly*, *18*(2)¸ pp. 82-88.

Ginther, A., & Grant, L. (1997). The influence of proficiency, language background, and topic on the production of grammatical form and error on the Test of Written English. In A. Huhta, V. Kohonen, L. Kurki-Suonio, & S. Luoma (Eds.), *Current developments and alternatives in language assessment: Proceedings of LTRC 1996* (pp. 385-397). Jyvaskyla: University of Jyvaskyla and University of Tampere.

Grabe, W., & Kaplan, R. B. (1996). *Theory and practice of writing*. New York: Longman.

Grabowski, J. (1996). Writing and speaking: Common grounds and differences toward a regulation theory of written language production. In C. M. Levy & S. Ransdell (Eds.), *The science of writing* (pp. 73-91). NJ: Lawrence Erlbaum Associates.

Greenberg, K. (1981). *The effects of variations in essay questions on the writing performance of CUNY freshmen.* New York: The City University of New York Instructional Resource Center.

Hake, R. (1986). How do we judge what they write? In K. L. Greenberg, H. S. Wiener, & R. A. Donovan (Eds.), *Writing assessment: Issues and strategies* (pp. 153-167). New York: Longman.

Hambleton, R. K. (1989). Principles and selected applications of item response theory. In R. L. Linn (Ed.), *Educational measurement,* (3rd ed., pp. 147-200). New York: American Council on Education/Macmillan.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory*. Newbury Park, CA: Sage.

Hamp-Lyons, L. (1989). Raters respond to rhetoric in writing. In H. W. Dechert & M. Raupach (Eds.), *Interlingual processes* (pp. 229-244). Tubingen, Germany: Gunter Narr Verlag.

Hamp-Lyons, L. (1990). Second language writing: Assessment issues. In B. Kroll (Ed.), *Second language writing: Research insights for the classroom*. Cambridge: Cambridge University Press.

Hamp-Lyons, L. (1991). Basic concepts. In L. Hamp-Lyons (Ed.), *Assessing second language writing in academic contexts* (pp. 5-15). Norwood, NJ: Ablex.

Hamp-Lyons, L., & Davies, A. (2008). The Englishes of English tests: Bias revisited. *World Englishes*, *27*(1), pp. 26-39.

Hamp-Lyons, L., & Mathias, S. P. (1994). Examining expert judgments of task difficulty on essay tests. *Journal of Second Language Writing*, *3*(1), pp. 49-68.

Han, Z. H. (2004). *Fossilization in Adult Second Language Acquisition*. Clevedon, UK: Multilingual Matters.

Henning, G. (1992). *Scalar analysis of the Test of Written English*. TOEFL Research Reports, RR-92-30. Princeton, New Jersey: Educational Testing Service.

Henning, G. (1996). Accounting for nonsystematic error in performance ratings. *Language Testing*, *13* (1), pp. 53-61.

Hill, K. (1996). Who should be the judge? The use of non-native speakers as raters on a test of English as an international language. *Melbourne Papers in Language Testing*, *5*(2), pp. 29-50.

Hinkel, E. (2002). *Second language writers' text: Linguistic and rhetorical features.* Mahwah, NJ: Lawrence Erlbaum.

Hoetker, J. (1982). Essay examination topics and students' writing. *College Composition and Communication, 33*(4), pp. 377-392.

Hoetker, J., & Brossell, G. (1989). The effects of systematic variations in essay topics on the writing performance of college freshmen. *College Composition and Communication, 40*(4), pp. 414-421.

Homburg, T. J. (1984). Holistic evaluation of ESL compositions: Can it be validated objectively? *TESOL Quarterly*, *18*(1), pp. 87-107.

Horowitz, D. (1991). ESL writing assessments: Contradictions and resolutions. In L. Hamp-Lyons (Ed.), *Assessing second language writing in academic contexts* (pp. 71-86). Norwood, NJ: Ablex.

Hoyt, W. T., & Kerns, M. D. (1999). Magnitude and moderators of bias in observer ratings: A meta-analysis. *Psychological Methods*, *4*(4), pp. 403-424.

Hughes, A. (1989). *Testing for language teachers*. Cambridge: Cambridge University Press.

Huot, B. A. (1990). The literature of direct writing assessment: Major concerns and prevailing trends. *Review of Educational Research*, *60*(2), pp. 237-263.

Huot, B. A. (1993). The influence of holistic scoring procedures on reading and rating student essays. In M. M. Williamson & B. A. Huot (Eds.), *Validating holistic scoring for writing assessment* (pp. 206-232). Cresskill, NJ: Hampton Press.

Hyland, K. (2002). *Teaching and researching writing*. Harlow, England: Longman.

Hymes, D. H. (1967). Models of the interaction of language and social setting. *Journal of Social Issues*, *23*(2), pp. 8-38.

Hymes, D. H. (1972). On communicative competence. In J. J. Gumperz & D. Hymes. *Directions in sociolinguistics: the ethnography of communication*. New York: Holt Rinehart and Winston.

IELTS (2008). *IELTS: International English Language Testing System*. Retrieved February 15, 2008, from http://www.ielts.org

Janopoulous, M. (1992). University faculty tolerance of NS and NNS writing errors: A comparison. *Journal of Second Language Writing*, *1*(2), pp. 109-121.

Jennings, M., Fox, J., Graves, B., & Shohamy, E. (1999). The test-takers’ choice: An investigation of the effect of topic on language-test performance. *Language Testing*, *16*(4), pp. 426-456.

Johnson, J. S. (2004). *Score gains for repeat MELAB administrations*. ELI Research Reports 2004-04. Ann Arbor, MI: English Language Institute, University of Michigan.

Johnson, J. S. (2005). *MELAB 2004 descriptive statistics and reliability estimates*. ELI Research Reports 2005-03. Ann Arbor, MI: English Language Institute, University of Michigan.

Johnson, J. S. (2006). *MELAB 2005 descriptive statistics and reliability estimates*. ELI Research Reports 2006-02. Ann Arbor, MI: English Language Institute, University of Michigan.

Johnson, J. S. (2007). *MELAB 2006 descriptive statistics and reliability estimates*. ELI Research Reports 2007-03. Ann Arbor, MI: English Language Institute, University of Michigan.

Johnson, J. S., & Lim, G. S. (2009). The influence of rater language background on writing performance assessment. *Language Testing*, *26*(4), pp. 1-21.

Johnson, J. S., & Song, T. (2008). *MELAB 2007 descriptive statistics and reliability estimates.* ELI Research Reports 2008-03. Ann Arbor, MI: English Language Institute, University of Michigan.

Kachru, B. B. (Ed.) (1992). *The Other Tongue: English across cultures*. 2nd edition. Urbana: University of Illinois Press.

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

Kane, M. T. (2001). Current concerns in validity theory. *Journal of Educational Measurement*, *38*(4), pp. 319-342.

Kane, M. T. (2002). Validating high-stakes testing programs. *Educational Measurement: Issues and Practice*, *21*(1), pp. 31-41.

Kane, M. T. (2004). Certification testing as an illustration of argument-based validation. *Measurement: Interdisciplinary Research and Perspectives*, *2*(3), pp. 135-170.

Kane, M. T. (2006). Validation. In R. L. Brennan (Ed.), *Educational Measurement*, (4th ed., pp. 17-64). Westport, CT: American Council on Education.

Kane, M. T., Crooks, T., & Cohen, A. (1999). Validating measures of performance. *Educational Measurement: Issues and Practice*, *18*(2), pp. 5-17.

Kaplan, R. B. (1966). Cultural thought patterns in intercultural education. *Language Learning*, *16*(1), pp. 1-20.

Kim, M. K. (2001). Detecting DIF across the different language groups in a speaking test. *Language Testing*, *18*(1), pp. 89-114.

Kim, Y. H. (2009). An investigation into native and non-native teachers' judgments of oral English performance: A mixed methods approach. *Language Testing*, *26*(2), pp. 187-217.

Knoch, U. (2009, March). Investigating the effectiveness of individualized feedback to rating behavior on a longitudinal study. Paper presented at the 30th Language Testing Research Colloquium, Denver, Colorado.

Kobayashi, T. (1992). Native and nonnative reactions to ESL compositions. *TESOL Quarterly*, *26*(1), pp. 81-112.

Kobayashi, H. & Rinnert, C. (1996). Factors affecting composition evaluation in an EFL context: Cultural rhetorical pattern and readers' background. *Language Learning*, *46*(3), pp. 397-437.

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

Kozaki, Y. (2004). Using GENOVA and FACETS to set multiple standards on performance assessment for certification in medical translation from Japanese to English. *Language Testing*, *21*(1), pp. 1-27.

Kroll, B. (1998). Assessing writing abilities. *Annual Review of Applied Linguistics*, *18*, pp. 219-240.

Kroll, B., & Reid, J. (1994). Guidelines for designing writing prompts: Clarifications, caveats, and cautions. *Journal of Second Language Writing*, *3*(3), pp. 231-255.

Kunnan, A. J. (Ed.) (2000). *Fairness and validation in language assessment: Selected papers from the 19th Language Testing Research Colloquium, Orlando, Florida.* Cambridge: Cambridge University Press.

Lado, R. (1961). *Language Testing*. New York: McGraw-Hill.

Lane, S., & Stone, C. A. (2006). Performance assessment. In R. L. Brennan (Ed.), *Educational measurement*, (4th ed., pp. 387-431). Westport, CT: American Council on Education.

Larsen-Freeman, D., & Cameron, L. (2008). *Complex systems and applied linguistics*. Oxford: Oxford University Press.

Lee, H. K., & Anderson, C. (2007). Validity and topic generality of a writing performance test. *Language Testing*, *24*(3), pp. 307-330.

Lee, Y. W., Breland, H., & Muraki, E. (2004). *Comparability of TOEFL CBT prompts for different native language groups*. TOEFL Research Reports, RR-04-24. Princeton, NJ: Educational Testing Service.

Leki, I. (1991). Twenty-five years of contrastive rhetoric: Text analysis and writing pedagogues. *TESOL Quarterly*, *25*(1), pp. 123-139.

Lewkowicz, J. (1997). Investigating authenticity in language testing. Unpublished doctoral dissertation, University of Lancaster.

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

Linacre, J. M. (1997). Guidelines for rating scales. *MESA Research Note #2.* Retrieved June 24, 2009, from http://www.rasch.org/rn2.htm

Linacre, J. M. (2002). What do infit, outfit, mean-square and standardized mean? *Rasch Measurement Transactions, 16*, p. 878.

Linacre, J. M. (2006). *Facets Rasch measurement computer program*. Chicago: Winsteps.com.

Linn, R. L., Baker, E. L., & Dunbar, S. B. (1991). Complex, performance-based assessment: Expectations and validation criteria. *Educational Researcher*, *20*(2), pp. 15-21.

Long, M. H. (1985). A role for instruction in second language acquisition: Task-based language teaching. In K. Hyltenstam & M. Pienemann (Eds.), *Modelling and assessing second language acquisition* (pp. 77-99). Clevedon, England: Multilingual Matters.

Lowenberg, P. H. (2000). Non-native varieties and issues of fairness in testing English as a world language. In A. J. Kunnan (Ed.), *Fairness and validation in language assessment: Selected papers from the 19th Language Testing Research Colloquium, Orlando, Florida* (pp. 43-59). Cambridge: Cambridge University Press.

Lumley, T. (2002). Assessment criteria in a large-scale writing test: what do they really mean to raters? *Language Testing*, *19*(3), pp. 246-276.

Lumley, T. (2006). *Assessing second language writing: The rater's perspective*. Frankfurt am Main: Peter Lang.

Lumley, T., & McNamara, T. F. (1995). Rater characteristics and rater bias: Implications for training. *Language Testing*, *12*(1), 54-71.

Lumley, T., & O'Sullivan, B. (2005). The effect of test-taker gender, audience and topic on task performance in tape-mediated assessment of speaking. *Language Testing*, *22*(4), pp. 415-437.

Lunz, M. E., & Stahl, J. A. (1990). Judge consistency and severity across grading periods. *Evaluation and the Health Professions*, *13*(4), pp. 425-444.

Lunz, M. E., & Stahl, J. A. (1993). The effect of rater severity on person ability measures: A Rasch model analysis. *American Journal of Occupational Therapy*, *47*(4), pp. 311-317.

McMillan, J. H., & Schumacher, S. (2001). *Research in education: A conceptual introduction* (5th Ed.). New York: Longman.

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

McNamara. T. F., & Roever, C. (2006). *Language testing: The social dimension*. Language Learning Monograph Series. Malden, MA: Blackwell Publishing.

Messick, S. (1989). Validity. In R. Linn (Ed.), *Educational Measurement*. New York: Macmillan.

Messick, S. (1994). The interplay of evidence and consequences in the validation of performance assessments. *Educational Researcher*, *23*(2), pp. 13-23.

Messick, S. (1996). Validity and washback in language testing. *Language Testing*, *13*(3), pp. 241-256.

Milanovic, M., Saville, N., & Shen, S. (1996). A study of the decision-making behaviour of composition markers. In M. Milanovic & N. Saville (Eds.), *Performance testing, cognition, and assessment. Selected papers from the 15th Language Testing Research Colloquium, Cambridge and Arnhem* (pp. 92-114). Cambridge: Cambridge University Press and University of Cambridge Local Examinations Syndicate.

Miller, M. D., & Legg, S. M. (1993). Alternative assessment in a high-stakes environment. *Educational Researcher*, *12*(2), pp. 9-15.

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

Mohan, B., & Lo, W. A. Y. (1985). Academic writing and Chinese students: Transfer and developmental factors. *TESOL Quarterly*,*19*(3), pp. 515-534.

Moore, T., & Morton, J. (1999). Authenticity in the IELTS academic module writing test: A comparative study of Task 2 items and university assignments. In R. Tulloh (Ed.), *IELTS research reports*, *volume 2* (pp. 64-106). Canberra: IELTS Australia.

Morrow, K. (1979). Communicative language testing: revolution or evolution? In C. J. Brumfit & K. Johnson (Eds.), *The communicative approach to language teaching*. Oxford: Oxford University Press.

Moss, P. (1992). Shifting conceptions of validity in educational measurement: Implications for performance assessment. *Review of Educational Research*, *62*(3), pp. 229-258.

Munby, J. (1978). *Communicative syllabus design*. Cambridge: Cambridge University Press.

Nisbett, R. E. (2003). *The geography of thought: How Asians and Westerners think differently…and why*. New York: The Free Press.

Norris, J., Brown, J.D., Hudson, T., & Yoshioka, J. (1998). *Designing second language performance assessments*. Technical Report No. 18. Hawaii: University of Hawaii Press.

O'Hagan, S. (1999). Assessment criteria for non-native speaker and native speaker essays: Do uniform standards work? *Melbourne Papers in Language Testing*, *8*(2), pp. 20-52.

O'Loughlin, K. (1992). Do English and ESL teachers rate essays differently? *Melbourne Papers in Language Testing*, *1*(2), pp. 19-44.

O'Loughlin, K., & Wigglesworth, G. (2007). Investigating task design in academic writing prompts. In L. Taylor & P. Falvey (Eds.), *IELTS collected papers. Research in speaking and writing performance* (pp. 379-421). Cambridge: Cambridge University Press.

O'Neill, T. R., & Lunz, M. E. (1996, April). Examining the invariance of rater and project calibrations using a multi-facet Rasch model. Paper presented at the annual meeting of the American Educational Research Association, New York, NY.

O'Sullivan, B., & Rignall, M. (2007). Assessing the value of bias analysis feedback to raters for the IELTS writing module. In L. Taylor & P. Falvey (Eds.), *IELTS collected papers. Research in speaking and writing performance* (pp. 446-478). Cambridge: Cambridge University Press.

Oxford English dictionary (2nd ed.). (1996). Oxford: Oxford University Press.

Park, T. J. (2006). Detecting DIF across different language and gender groups in the MELAB essay test using the logistic regression method. *Spaan Fellow Working Papers in Second or Foreign Language Assessment*, *4*, pp. 81-94.

Park, Y. M. (1988). Academic and ethnic background as factors affecting writing performance. In A. C. Purves (Ed.), *Writing across languages and cultures: Issues in contrastive rhetoric* (pp. 261-272). Newbury Park, CA: SAGE Publications.

Penfield, R. D., & Lam, T. C. M. (2000). Assessing differential item functioning in performance assessment: Review and recommendations. *Educational Measurement: Issues and Practice*, *19*(3), pp. 5-15.

Polio, C., & Glew, M. (1996). ESL writing assessment prompts: How students choose. *Journal of Second Language Writing*, *5*(1), pp. 35-49.

Pollitt, A., & Hutchinson, C. (1987). Calibrating graded assessments: Rasch partial credit analysis of performance in writing. *Language Testing*, *4*(1), pp. 72-92.

Pollitt, A., & Murray, N. L. (1996). What raters really pay attention to. In M. Milanovic & N. Saville (Eds.), *Performance testing, cognition, and assessment. Selected papers from the 15th Language Testing Research Colloquium, Cambridge and Arnhem* (pp. 74-91). Cambridge: Cambridge University Press and University of Cambridge Local Examinations Syndicate.

Powers, D. E., & Fowles, M. E. (1998). *Test takers' judgments about GRE writing test prompts*. ETS Research Report 98-36. Princeton, NJ: Educational Testing Service.

Powers, D. E., Fowles, M. E., Farnum, M., & Gerritz, K. (1992). *Giving a choice of topics on a test of basic writing skills: Does it make any difference?* ETS Research Report No. 92-19. Princeton, NJ: Educational Testing Service.

Pula, J. J., & Huot, B. A. (1993). A model of background influences on holistic raters. In M. M. Williamson & B. A. Huot (Eds.), *Validating holistic scoring for writing assessment: Theoretical and empirical foundations* (pp. 237-265). Cresskill, NJ: Hampton Press.

Purpura, J. E. (2005). Michigan English language assessment battery (MELAB). In S. Stoynoff & C. A. Chapelle (Eds.), *ESOL tests and testing* (pp. 87-91). Alexandria, Virginia: TESOL.

Purves, A. C. (1992). Reflections on research and assessment in written composition. *Research in the Teaching of English*, *26*(1), pp. 108-122.

Quellmalz, E. S., Capell, F. J., & Chou, C. P. (1982). Effects of discourse and response mode on the measurement of writing competence. *Journal of Educational Measurement*, *19*(4), pp. 241-258.

Rasch, G. (1980). *Probabilistic models for some intelligence and attainment tests*. Chicago: University of Chicago Press.

Reed, D. J. & Cohen, A. D. (2001). Revisiting raters and ratings in oral language assessment. In C. Elder, A. Brown, E. Grove, K. Hill, N. Iwashita, T. Lumley, T. McNamara, & K. O'Loughlin (Eds.), *Experimenting with uncertainty: Language testing essays in honor of Alan Davies* (pp. 82-96). Cambridge: Cambridge University Press.

Reid, J. (1990). Responding to different topic types: A quantitative analysis from a contrastive rhetoric perspective. In B. Kroll (Ed.), *Second language writing: Research insights for the classroom* (pp. 191-209). Cambridge: Cambridge University Press.

Rinnert, C., & Kobayashi, H. (2001). Differing perceptions of EFL writing among readers in Japan. *Modern Language Journal, 85*(2), pp. 189-209.

Robinson, P. (1995). Task complexity and second language narrative discourse. *Language Learning, 45*(1), pp. 99-140.

Ruth, L., & Murphy, S. (1988). *Designing writing tasks for the assessment of writing*. Norwood, NJ: Ablex.

Saal, F. E., Downey, R. G., & Lahey, M. A. (1980). Rating the ratings: Assessing the psychometric quality of rating data. *Psychological Bulletin, 88*(2), pp. 413-428.

Sakyi, A. A. (2000). Validation of holistic scoring for ESL writing assessment: How raters evaluate compositions. In A. J. Kunnan (Ed.), *Fairness and validation in language assessment: Selected papers from the 19th Language Testing Research Colloquium, Orlando, Florida* (pp. 129-152). Cambridge: Cambridge University Press.

Santos, T. (1988). Professors' reactions to the academic writing of nonnative-speaking students. *TESOL Quarterly*, *22*(1), 69-90.

Schaefer, E. (2008). Rater bias patterns in an EFL writing assessment. *Language Testing, 25*(4), pp. 465-493.

Schaeffer, G. A., Briel, J. B., & Fowles, M. E. (2001). *Psychometric evaluation of the new GRE writing assessment.* GRE Board Report, 96-11. Princeton, NJ: Educational Testing Service.

Schoonen, R., Vergeer, M., & Eiting, M. (1997). The assessment of writing ability: Expert readers versus lay readers. *Language Testing*, *14*(2), pp. 157-184.

Selinker, L. (1972). Interlanguage. *International Review of Applied Linguistics*, *10*(2), pp. 209-231.

Shavelson, R. J., Baxter, G. P., & Gao, X. (1993). Sampling variability of performance assessments. *Journal of Educational Measurement*, *30*(3), pp. 215-232.

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

Shaw, S. D., & Weir, C. J. (2007). *Examining writing: Research and practice in assessing second language writing*. Cambridge: Cambridge University Press.

Shi, L. (2001). Native- and nonnative-speaking EFL teachers' evaluation of Chinese students' English writing. *Language Testing*, *18*(3), pp. 303-325.

Shohamy, E. (2001). *The power of tests: A critical perspective on the uses of language tests*. London: Pearson.

Shohamy, E., Gordon, C. M., & Kraemer, R. (1992). The effect of raters' background and training on the reliability of direct writing tests. *The Modern Language Journal*, *76*(1), pp. 27-33.

Silva, T. (1993). Toward an understanding of the distinct nature of L2 writing: The ESL research and its implications. *TESOL Quarterly*, *27*(4), pp. 657-677.

Skehan, P. (1996). A framework for the implementation of task-based instruction. *Applied Linguistics*, *17*(1), pp. 38-62.

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

Song, B., & Caruso, I. (1996). Do English and ESL faculty differ in evaluating the essays of native English-speaking and ESL students? *Journal of Second Language Writing*, *5*(2), pp. 163-182.

Spaan, M. (1993). The Effect of Prompt in Essay Examinations. In D. Douglas & C. Chapelle (Eds.), *A new decade of language testing research* (pp. 98-122). Alexandria, VA: TESOL.

Spolsky, B. (Ed.) (1978). *Approaches to language testing*. *Advances in language testing series: 2*. Arlington, VA: Center for Applied Linguistics.

Spolsky, B. (1995). *Measured words*. Oxford: Oxford University Press.

Stock, P. L., & Robinson, J. L. (1987). Taking on testing. *English Education*, *19*(1), pp. 93-121.

Swaminathan, H., & Rogers, H. J. (1990). Detecting differential item functioning using logistic regression procedures. *Journal of Educational Measurement*, *27*(4), pp. 361-370.

Tedick, D. J. (1990). ESL writing assessment: Subject-matter knowledge and its impact on performance. *English for Specific Purposes*, *9*(2), pp. 123-143.

Tedick, D. J., & Mathison, M. (1995). Holistic scoring in ESL writing assessment: What does an analysis of rhetorical features reveal? In D. Belcher & G. Braine (Eds.), *Academic writing in a second language: Essays on research and pedagogy* (pp. 205-230). Norwood, NJ: Ablex.

Toulmin, S. E. (1958). *The use of argument*. Cambridge: Cambridge University Press.

Van Ek, J. (1975). *Systems development in adult language learning: The threshold level in a European unit credit system for modern language learning by adults*. Strasbourg: Council of Europe.

Vaughan, C. (1991). Holistic assessment: What goes on in the rater's mind? In L. Hamp-Lyons (Ed.), *Assessing second language writing in academic contexts*, (pp. 111-126). Norwood, NJ: Ablex.

Wainer, H., & Thissen, D. (1994). On examinee choice in educational testing. *Review of Educational Research, 64*(1), pp. 159-195.

Wainer, H., Wang, X. B., & Thissen, D. (1994). How well can we compare scores on test forms that are constructed by examinees' choice? *Journal of Educational Measurement*, *31*(3), pp. 183-199.

Wang, X. B., Wainer, H., & Thissen, D. (1995). On the viability of some untestable assumptions in equating exams that allow examinee choice. *Applied Measurement in Education*, *8*(3), pp. 211-225.

Weigle, S. C. (1994). Effects of training on raters of ESL compositions. *Language Testing*, *11*(2), pp. 197-223.

Weigle, S. C. (1998). Using FACETS to model rater training effects. *Language Testing*, *15*(2), pp. 263-87.

Weigle, S. C. (1999). Investigating rater/prompt interactions in writing assessment: Quantitative and qualitative approaches. *Assessing Writing*, *6*(2), pp. 145-178.

Weigle, S. C. (2000). Test review: The Michigan English language assessment battery (MELAB). *Language Testing*, *17*(4), pp. 449-455.

Weigle, S. C. (2002). *Assessing Writing*. Cambridge: Cambridge University Press.

Weigle, S. C., Bolt, H., & Valsecchi, M. I. (2003). Effects of task and rater background on the evaluation of ESL student writing: A pilot study. *TESOL Quarterly*, *37*(2), pp. 345-354.

Weir, C. J. (2003). A survey of the history of the Certificate of Proficiency in English (CPE) in the twentieth century. In C. J. Weir & M. Milanovic (Eds.), *Continuity and innovation: The history of the CPE 1913-2002* (pp. 1-56). Cambridge: Cambridge University Press.

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

Widdowson, H. G. (1978). *Teaching language as communication.* Oxford: Oxford University Press.

Wigglesworth, G. (2007). Task and performance based assessment. In N. H. Hornberger (Series Ed.) & E. Shohamy & N. H. Hornberger (Vol. Eds.), *Encyclopedia of language and education: Vol. 7*, (2nd ed., pp. 111-122). Boston, MA: Springer.

Willingham, H. H., & Cole, N. S. (1997). *Gender and fair assessment*. Mahwah, NJ: Lawrence Erlbaum.

Wiseman, C. S. (2009, March). *Rater decision-making behaviors in measuring second language writing ability using holistic and analytic scoring methods*. Paper presented at the annual meeting of the American Association for Applied Linguistics, Denver, Colorado.

Wolfe, E. W. (1997). The relationship between essay reading style and scoring proficiency in a psychometric scoring system. *Assessing Writing*, *4*(1), pp. 83-106.

Wolfe, E. W., Kao, C. W., & Ranney, M. (1998). Cognitive differences in proficient and nonproficient essay scorers. *Written Communication*, *15*(4), pp. 465-492.

Wright, B. D. (1984). Despair and hope for educational measurement. *Contemporary Education Review*, *3*(1), pp. 281-288.

Wright, B. D., Linacre, M., Gustafsson, J. E., & Martin-Loff, P. (1994). Reasonable mean-square fit values. *Rasch Measurement Transactions*, *8*(3), p. 370.

Wright, B. D., & Masters, G. N. (1981). *The measurement of knowledge and attitude*. Research Memorandum No. 30. Chicago: University of Chicago, MESA Psychometric Laboratory.

Xi, X. (2007). Methods of test validation. In N. H. Hornberger (Series Ed.) & E. Shohamy & N. H. Hornberger (Vol. Eds.), *Encyclopedia of language and education: Vol. 7*, (2nd ed., pp. 177-196). Boston, MA: Springer.

Yu, G. (2007). Lexical diversity in MELAB writing and speaking task performances. *Spaan Fellow Working Papers in Second or Foreign Language Assessment*, *5*, pp. 79-116.

Zhang, W. (1998). *The rhetorical patterns found in Chinese EFL student writers' examination essays in English and the influence of these patterns on rater response*. Unpublished doctoral dissertation, Hong Kong Polytechnic University.

Zumbo, B. D. (1999). *A handbook on the theory and methods of differential item functioning (DIF): Logistic regression modeling as a unitary framework for binary and Likert-type (ordinal) item scores.* Ottawa, Ontario, Canada: Directorate of Human Resources Research and Evaluation, Department of National Defense.