Spaan Fellow Working Papers in Second or Foreign Language Assessment

Volume 8: 95–116

English Language Institute
University of Michigan
www.lsa.umich.edu/eli/research/spaan

Investigating Prompt Effects in Writing Performance Assessment

Gad S. Lim
Ateneo de Manila University
University of Michigan

ABSTRACT Performance assessments have become the norm for evaluating language learners' writing abilities in international examinations of English proficiency. In these assessments, prompts are systematically varied for different test-takers, raising the possibility of a prompt effect and affecting the validity, reliability, and fairness of these tests. This study uses data from the Michigan English Language Assessment Battery (MELAB), covering a period of over four years (n ratings = 29,831), to examine this issue. It uses the multi-facet extension of Rasch methodology to investigate the comparability of prompts that differ on topic domain, rhetorical task, prompt length, task constraint, expected grammatical person of response, and number of tasks. It also considers whether prompts are differentially difficult for test takers of different genders, language backgrounds, and proficiency levels. The results show that, on the whole, test-takers' scores reflect ability in the construct being measured and are generally not affected by a range of prompt dimensions, or test taker characteristics. It can be concluded that scores on this test and others whose particulars are like it have score validity.

Introduction

In international examinations of English language proficiency, performance assessment has become the norm in assessing the productive skills. 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 are typically judged by human raters (American Educational Research Association, American Psychological Association, and National Council on Measurement in Education, 1999; Kane, Crooks, & Cohen, 1999; McNamara, 1996). In assessments of second language writing, performance assessment has taken the modal form of the timed, impromptu writing test (Weigle, 2002). The use of performance assessment is in keeping with communicative approaches and conceptions of language ability, and compared to discrete item and indirect tests, these tests are seen as possessing greater theoretical and construct validity (Kane, et al., 1999; Linn, Baker, & Dunbar, 1991; Moss, 1992). In addition, they are thought to have the added value of providing positive washback (Miller & Legg, 1993).

However, there are also challenges associated with the use of performance assessments. Because performance assessments tend to require more time, examinees are typically tested on one or two tasks and evaluated on the basis of these limited samples. It is unclear whether performance on a small number of tasks is sufficient for representing a domain as apparently complex and multi-faceted as writing ability. That is, there is the risk of construct underrepresentation (Messick, 1989, 1994, 1996). Additionally, test takers are usually given one or two prompts from a larger pool of prompts. 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. How comparable are the performances of a test taker who responds to one prompt and another test taker who responds to another prompt? In other words, there is also the risk of construct-irrelevant variance (Messick, 1989, 1994, 1996) or what Jennings, Fox, Graves, and Shohamy (1992) have called a "prompt effect." These issues do not just raise questions about validity and reliability; perhaps more importantly, they raise questions of fairness (Kunnan, 2000), which examination providers must address.

This study aims to address those questions to a certain extent. Taking a look at one exam of English language proficiency—the Michigan English Language Assessment Battery (MELAB)—it investigates those characteristics that might contribute to prompts not being comparable, and determine whether prompt effects indeed exist.

Literature Review

As in all language use, responding to prompts requires topic knowledge. Where prompts are concerned, the usual approach of language proficiency exams is to use topics that all test takers are expected to know, and perhaps to give them a small selection of such topics (Bachman & Palmer, 1996). However, the question of relative prompt difficulty remains, and what makes a prompt easy or difficult still eludes people, test takers and test makers alike (cf. Chiste & O'Shea, 1988; Dobson, Spaan, & Yamashiro, 2003; Freedman, 1983; Hamp-Lyons & Mathias, 1994; Power & Fowles, 1998). A number of features (e.g. subject matter, rhetorical specification) have been identified that possibly contribute to prompts being easier or more difficult. Test taker characteristics such as gender and language background have also been identified that may interact with these features. These are now discussed.

Subject Matter

First is subject matter or topic domain. While the topics used in exams are presumed to be familiar to all test takers, it remains that some test takers may 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. These were also the reasons cited for choosing a topic 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. Test takers in Powers and Fowles (1998) did no better on topics they preferred. When the English Language Testing System was being revised, the plan to divide test takers into six discipline areas was abandoned when it was found that there were

no systematic differences in test-takers' performances when responding to general and field-specific prompts (Hamp-Lyons, 1990). On the other hand, Tedick (1990) reports that ESL graduate students did better on topics specific to their field than on general topics. The prompts used in the study might be worth looking into, however. The general prompt is provided first, 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)

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. However, the topic is constrained in that one can only write about progress and laziness and nothing else. The "specific" prompt, ironically, is the more general prompt. The field-specific prompt is virtually unconstrained, leaving respondents plenty of leeway in choosing what to write about. That the topic is controversial 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. From this study, a possible prompt factor emerges then: those that allow one to respond in a 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. Two possible prompt-related factors have been identified here. One is topic domain; the other, task constraint.

Rhetorical Task

Studies on the type of writing called for in a prompt have by and large compared personal versus impersonal writing, or narrative versus argumentative writing. A number of studies have 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; Hoetker & Brossell, 1989). These studies found no significant differences, though this lack of finding can perhaps be attributed to the cues 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.

It can be seen from the above that the difference between the two prompts are difficult to spot. However, in the case of Hoetker and Brossell (1989), 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.

Other studies have focused on rhetorical task (Hamp-Lyons & Mathias, 1994; Hinkel, 2002; Quellmalz, Capell, & Chou, 1982; Spaan, 1993; Wiseman, 2009). These studies have found, contrary to the expectations of experts, that test-takers did better on argumentative tasks than on narrative tasks. Quellmalz, et al. (1982), 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 (2009) looked at a college writing placement test and had the same findings. Similarly, Hamp-Lyons and Mathias (1994) found that argumentative/public compositions were scored higher than expository (narrative/descriptive)/private compositions in their sample of MELAB test takers. The one exception to these is Spaan (1993), who 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.

Task Specification

The way prompts are specified has received some amount of attention. 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 prompt, framed prompt, and text-based or reading based prompt. 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 interpret, react to, or apply the information in those readings. For his part, Brossell (1983) divides 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 this difference in production 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; Lewkowicz, 1997).

Another approach to classifying prompt specification is by 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). 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).

Test-Taker Characteristics

Investigations of test-taker characteristics that could interact with prompt-related factors have focused on gender, language background, and proficiency level. Where test-taker gender is concerned, Breland, Bridgeman, and Fowles (1999), Breland, Lee, Najarian, and Muraki (2004), and Broer, Lee, Rizavi, and Powers (2005) have found instances of differential item functioning (DIF) in favor of female test takers in six different performance writing tests, to a magnitude up to 0.2 of a standard deviation. The authors caution though that the direction and size of the differences are highly sensitive to sample selection, and the findings should not be generalized beyond the exams studied.

Studies have also 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, or Spanish on a comparison and contrast task and on a graph/data commentary task. She found that writers from three of the language backgrounds, with the exception of the Spanish group, showed greater production on the graph task. There was also greater use of passive-voice 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 were found according to language background and area of academic specialization.

A number of the studies have also investigated the relationship between prompt and language background (Breland, et al., 1999; Broer, et al., 2005). The study by Breland, et al. compared ESL Hispanics and Asian Americans to White Americans, and found the prompts favoring the latter by 0.72 to 0.76 standard deviation units. The Broer, et al. study found a moderate-sized difference in favor of those whose strongest language was English. Finally, Lee, Breland, and Muraki (2004) compared test takers with Indo-European and East Asian first languages. That is, where the comparison groups in other studies have been people for whom English is a first language, this study compared two groups of non-native English writers. There were small uniform and non-uniform DIF for a minority of prompts, but 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. In general, taking language background as a factor, there is a notable difference in findings depending on the comparison group; DIF is more likely to show up when test takers for whom English is a first language are included.

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, 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.

Research Questions

In light of the literature, the following research questions can be asked:

1. 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?
2. To what extent are writing prompts not differentially difficult for test takers of different genders, language backgrounds, and proficiency level?

Method

The Test

The MELAB is an advanced-level English proficiency test for adults who use English as a second or foreign language, and who use the scores for various academic and professional purposes. The test includes sections assessing each of the four language skill areas. In the writing section, examinees are given 30 minutes to compose a handwritten composition on one of two prompts, which test takers do not see in advance. Each composition is scored using a holistic, 10-point scale by at least two raters. If the two ratings differ by more than one scale-point, a third rater adjudicates. The final score is the average of the ratings that are either equal or different by one scale-point (English Language Institute, 2005). Examinees 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.

The Prompts

The study's data includes 60 different prompts. They range in length from 12 to 82 words, with a mean of 38.47 (Table 1). In terms of sentences they were as short as a single sentence and as long as five sentences.

Table 1. Length of MELAB Writing Prompts

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

Unlike length, the other prompt dimensions that the study is concerned with—topic domain, rhetorical task, task constraint, expected grammatical person of response, and number of tasks—cannot be arrived at by mere counting. These dimensions were independently coded according to the categories in Table 2 by two testing professionals with expertise in writing assessment. The categories for topic domain are those used internally by the ELI, while the categories for the other dimensions came out of the literature.

Table 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, 2, 3…, n

After initial coding, the two coders met for a reconciliation meeting to agree on a common code in instances where they disagreed. They also chose to leave certain "disagreements" as they were, rather than force an agreement that might misrepresent the nature of those prompts. Their agreement rates before and after the meeting are given in Table 3.

Table 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

The Test Takers

The study's data include all test takers who took the MELAB between October 2003 and February 2008, and all the ratings assigned to their compositions, minus those with missing data. The resulting sample included 29,831 ratings for 10,536 test takers. 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 accounted 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 languages and language groups accounting for at least one percent of the total sample size are given in Table 4. (Language group refers to languages which have multiple dialects, e.g., Amoy, Cantonese, Hakka, and Mandarin were all coded under "Chinese").

Table 4. 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

It should be noted that there are a number of test takers whose first language is English, and for whom the test is not designed. Johnson and Lim (2009) 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.

Data Analysis

To analyze the data, this study employed multi-facet Rasch (Linacre, 1989; 2006), which models different elements of interest and puts them on a common, interval scale, thus facilitating meaningful comparisons among elements. The model can account for rater effects, thus providing accurate estimates for prompts. In addition, bias analysis can also be performed, thus making it ideal for this study's purposes.

In doing multi-facet Rasch analysis, it is important that the data be connected and that there be no "disjoint subsets". Earlier it was noted that the MELAB writing test asks test takers to choose between two prompts and to respond to just one. This creates a problem with connectedness. If each person responds to only one prompt, it is impossible to tell if any differences observed are due to the prompt or to some characteristic of those persons who were assigned/who chose that particular prompt.

The approach taken by other studies to solving this problem is by creating matching variables—usually some overall language ability variable based on test-takers' scores in other skill areas—and then matching different test takers according to their similarity in that regard (e.g., Breland, Lee, Najarian, & Muraki, 2004; Broer, Lee, Rizavi, & Powers, 2005; Lee,

Breland, & Muraki, 2004). This is arguably an imperfect solution, as it requires making certain assumptions regarding the relationship between writing and other skills. Additionally, identical overall scores can mask differing skill profiles.

The data used in this study permitted an approach that did not have to make such strong assumptions. The data include a large number of test takers who took the MELAB more than once. Thus, in this study, test takers were matched according to similarities in test scores *and* the fact that those being matched were in fact the same person. Elapsed time between test sittings provided an additional control; the less time between sittings, the less likely a person's ability has changed. Taken together, there can be greater confidence that matches being made are warranted. A procedure was followed that maximized stringency while minimizing matches required. In total, a modest total of only 214 matches were required for data connection to be achieved. Full details of the matching procedure can be found in Lim (2009).

The software FACETS (Linacre, 2006) was used to perform multi-facet Rasch analysis. To fit the requirements of the software, the ratings—which in the original ten point scale ranged from 53 to 97—were converted into a 0 to 9 scale, where 0 = 53 and 9 = 97. A model was specified which included the following facets: test-taker, gender, first language, proficiency level, prompt, and rater. Proficiency level was a dummy variable anchored to zero. Bias analysis was also requested for prompt and gender, prompt and language background, and prompt and proficiency level.

To answer the first research question, the comparability of prompts was evaluated, prima facie, by looking at the prompt measurement report, which provides a variety of statistics regarding the prompts, individually and as a whole. Then, the fair measure averages for all 60 prompts were entered into SPSS 16.0 for Windows, along with their codes for the six prompt dimensions being investigated. Cases where coders chose not to agree were excluded. 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. The results of the F-test and the associated p-values for each ANOVA were examined for significant outcomes. Where significant outcomes are found, Levene's test for homogeneity of variances and Tukey's HSD post hoc test was used to see which categories were significantly different from each other.

The bias analyses from FACETS were examined to answer the second research question. In the output, the chi-square test examines the null hypothesis that all the combinations (e.g., of particular prompt and particular gender) are equal in difficulty. If the null hypothesis had to be rejected, and interaction effects were indeed present, the results were examined for appropriately measured values that were also significant – that is, those with z-scores higher than |1.96| and infit mean square values within the acceptable range. The difference between observed and expected scale point averages for those combinations were then examined to find out the direction and magnitude of the bias.

Results and Discussion

To gauge the comparability of prompts, the difficulty parameters for the prompts are considered. The estimates are provided in Table 5, arranged in order of difficulty from the easiest to the most difficult. The separation index for this set of prompts was 5.85, with a

reliability of .97, indicating that the prompts can be reliably 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.

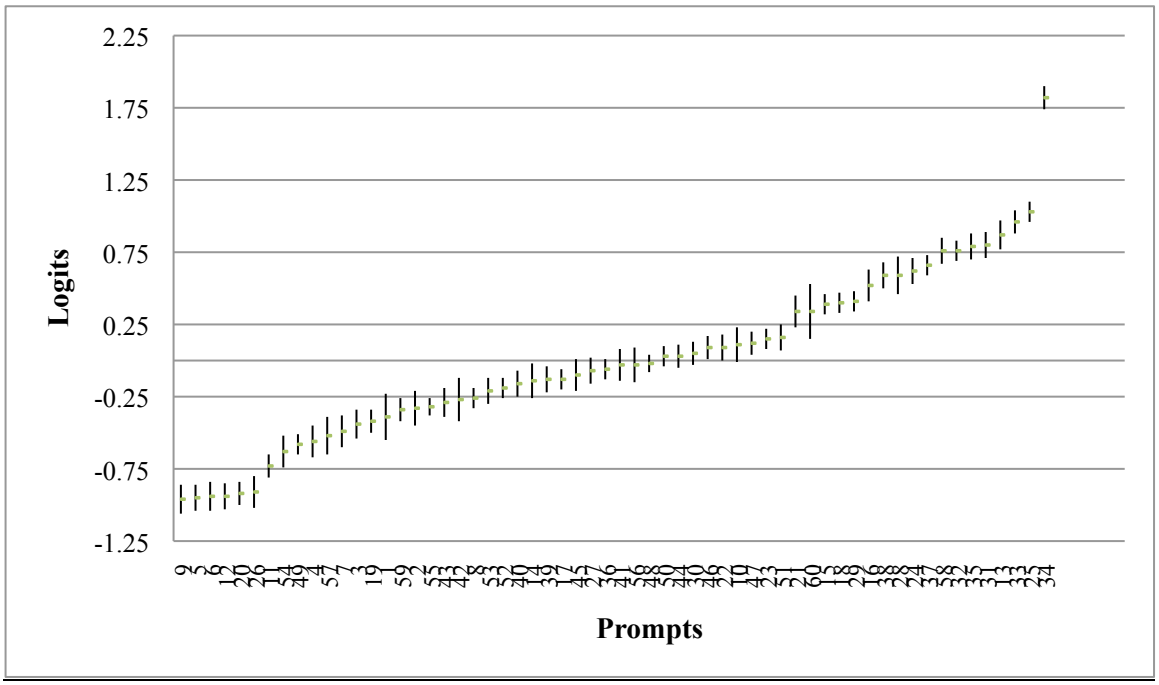

Figure 1. Range of Prompt Estimates, Arranged According to Severity

While the prompts significantly differ in difficulty, the real question is whether these significant differences are also meaningful. Figure 1 shows the difficulty measures of the prompts, accounting for standard error. It can be seen that Prompt 34 is a clear outlier, more than three standard deviations from the mean. The difficulty parameter of this prompt, allowing 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 1, 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. 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 the next most difficult prompts were excluded—say, Prompt 25 and 33—the range between the easiest and most difficult prompts would be further reduced to just 1.83 logits.

Table 5. Prompt Measurement Report

Prompt	n	Obsvd Ave.	Fair Ave.	Measure	S.E.	Infit MnSq	Infit 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
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

Assuming that these three prompts (25, 33, and 34) were excluded, what is the practical effect of the easiest and most difficult prompt differing by 1.83 logits? As was previously mentioned, multi-facet Rasch makes meaningful comparisons between different facets possible, as the rating scale has also been expressed in terms of the same logit scale. In the case of this analysis, the average range covered by each scale point is 3.88 logits. 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 the three outlier 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, the decision point for most MELAB users 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.

Research Question 1

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. Table 6 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 task constraint, grammatical person, and number of tasks.

Table 6. 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 7. 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 7. Prompt Dimensions Analyses of Variance

	df			
	Between Group	Within Group	F	Sig.
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 8). However, an inspection of the p-values indicated that the difference between business prompts and social prompts, 0.14 of a scale point, was approaching significance.

Table 8. 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

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.

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), and 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 longer the prompt, the lower the average score (Table 6). The only exception to this pattern was one sentence prompts. However, this relationship was not significant. It would appear, then, that reading a longer prompt might take somewhat more time, but not all that much, which accords with the findings of Polio & Glew (1996).

The one dimension that yielded significant differences was topic domain. Interestingly, in previous studies (Polio & Glew, 1996; Powers & Fowles, 1998) when asked what factors they considered in choosing prompts, test takers have overwhelmingly cited background knowledge and topic familiarity. Their intuition about what topic to choose 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.

Research Question 2

The second research question concerns the relationship between prompts and test-taker characteristics. Results of the bias/interaction analysis between prompt and gender, language background, and test-taker proficiency level are given in Tables 9, 10, and 11, respectively. Provided in the tables are the global measures, as well as individual interaction measures that are significant ($|z\text{-score}| > 1.96$). 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 9. Bias/Interaction Analysis: Prompt and Gender

Prompt x Gender	Obs-Exp Average	Bias+ Measure	Model S.E.	Z-Score	Infit MnSq	Outfit 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 10. Bias/Interaction Analysis: Prompt and Language Background

Prompt x Language Background	Obs-Exp Average	Bias+ Measure	Model S.E.	Z-Score	Infit MnSq	Outfit 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 11. Bias/Interaction Analysis: Prompt and Proficiency Level

Prompt x Proficiency Level	Obs-Exp Average	Bias+ Measure	Model S.E.	Z-Score	Infit MnSq	Outfit 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						

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. Note that the results for prompt and language proficiency do 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, which might contribute to the difference in findings. 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 for other variables 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.

Conclusion

The questions investigated by this study have to do with the fairness, validity, and reliability of second language writing performance assessments. The possible threat identified by the study is the systematic variation typically built into performance writing tests—in particular, that different test takers have to respond to different prompts, which may or may not be comparable in difficulty. 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.

The results of the 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 the 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.

References

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.

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.

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.

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

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.

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.

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.

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

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

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.

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.

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

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

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.

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., & Lim, G. S. (2009). The influence of rater language background on writing performance assessment. *Language Testing*, *26*(4), pp. 485–505.

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

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.

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.

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

Lim, G. S. (2009). Prompt and rater effects in second language writing performance assessment. Unpublished doctoral dissertation, University of Michigan.
Linacre, J. M. (1989). *Many-facet Rasch measurement*. Chicago: MESA Press.
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.
McNamara, T. F. (1996). *Measuring second language performance*. London: Longman.
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.
Miller, M. D., & Legg, S. M. (1993). Alternative assessment in a high-stakes environment. *Educational Researcher*, *12*(2), pp. 9–15.
Moss, P. (1992). Shifting conceptions of validity in educational measurement: Implications for performance assessment. *Review of Educational Research*, *62*(3), pp. 229–258.
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.
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.
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.
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.
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.
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.
Tedick, D. J. (1990). ESL writing assessment: Subject-matter knowledge and its impact on performance. *English for Specific Purposes*, *9*(2), pp. 123–143.

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

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.

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.