
Do Test Formats in Reading Comprehension Affect Second-Language Students' Test Performance Differently?

Ying Zheng, Liying Cheng, and Don A. Klinger

Large-scale testing in English affects second-language students not only greatly but also differently than first-language learners. The research literature reports that confounding factors in such large-scale testing such as varying test formats may differentially affect the performance of students from diverse backgrounds. An investigation of test performance between ESL/ELD students and non-ESL/ELD students on the Ontario Secondary School Literacy Test (OSSLT) was performed to investigate whether test formats in reading comprehension affected the two groups differently. The results indicate that the overall pattern of difficulty levels on the three test formats were the same between ESL/ELD students and non-ESL/ELD students, except that ESL/ELD students performed substantially lower on each format and that more variability was found among ESL/ELD students. Further, discriminant analysis results indicated that only the multiple-choice questions obtained a significant discriminant coefficient in differentiating the two groups. The results suggest a lack of association between test formats and test performance.

L'effet qu'ont les évaluations à grande échelle en anglais sur les élèves en langue seconde n'est pas seulement important, il est également différent de celui qu'elles ont sur les élèves en langue première. La recherche indique que dans les évaluations à grande échelle, les variables confusionnelles telles que les formats variés peuvent ne pas avoir le même effet sur des élèves d'origines différentes. La performance d'élèves en anglais langue seconde/développement de la langue anglaise (ESL/ELD) au test d'aptitude à lire et à écrire au secondaire de l'Ontario a été comparée à celle d'élèves qui n'étaient pas dans le programme ESL/ELD pour déterminer si le format de l'évaluation de la compréhension à l'écrit avait le même effet sur les deux groupes. Les résultats indiquent que la performance globale en termes de niveaux de difficulté aux trois formats de test était semblable pour les deux groupes. Toutefois, la performance des élèves ESL/ELD était sensiblement inférieure et plus variable pour chaque format. De plus, une analyse discriminante a révélé que seules les questions à choix multiples donnent une fonction discriminante significative lors de la comparaison des deux groupes. Les résultats donnent à penser qu'il n'y a pas de lien entre le format des tests et les performances des élèves.

Introduction

Research in language testing has pointed out that test-takers with different characteristics might be affected by a test in ways that are not relevant to the abilities being tested (Bachman, 1990; Kunnan, 1998). Test format has been shown to be an important facet that could influence different test-takers' test performance (Bachman & Palmer, 1982; Shohamy, 1984, 1997). The issue of test format differences has been the subject of debate because it is generally assumed that different test formats elicit different levels of skills or abilities; therefore, such tests are subject to having different effects on test-takers from various linguistic and cultural backgrounds. Kunnan (2004) raised the issue of test fairness, arguing that certain test formats may favor some groups of test-takers but not others, threatening the validity of a particular test. Shohamy (1997) claimed that language tests employing test methods that are unfair to different groups of test-takers are unethical. If group performance differences do exist, the reason should be real differences in the skills or abilities being tested instead of confounding variables such as test formats (Elder, 1997).

The present study aimed to determine whether test formats in reading comprehension on the Ontario Secondary School Literacy Test (OSSLT) affected English as a Second Language (ESL) and English Literacy Development (ELD) students differently than their non-ESL/ELD counterparts. ESL students are defined in the Ontario curriculum as students whose first language is not English, but who have received educational experience in their own countries using their first language. ELD students are those who are from countries or regions where access to education may have been limited and who have had few opportunities to develop literacy skills in any language (Ministry of Education and Training, 1999). ESL and ELD students are students who are identified by their school as ESL/ELD learners and who are recommended to take ESL and/or ELD courses. These students are also referred to as second-language students in this article. Unfortunately, information about the length of time these students had been in Canada, the level of their English proficiency, and any previous training experience for the OSSLT test or similar tasks was not available for this study.

Research Background

The OSSLT is a provincially mandated standardized test of English literacy. It is a graduation requirement for all Ontario secondary students in order to receive their secondary school diploma. Administered by the Education Quality and Accountability Office (EQAO), this test is designed to assess the literacy skills that students are expected to have learned in all subjects by the end of grade 9 in Ontario (ages 15-16). The test consists of two major components: writing and reading. In the writing component four types of writing

task are included: a summary, a series of paragraphs expressing an opinion, a news report, and an information paragraph. In the reading component are 100 questions about 12 reading selections based on three types of texts: information (50%), consisting of explanation and opinion; graphic (25%), consisting of graphs, schedules and instructions; and narrative (25%), consisting of stories and dialogues. The students are expected to demonstrate the following three reading skills as required: understanding directly stated ideas and information; understanding indirectly stated ideas and information; and making connections between personal experiences and information in a reading selection (these terms are used in conformity with the EQAO terms). Finally, the comprehension questions employed to assess students' reading abilities are in three test formats: multiple-choice (MC) questions, constructed-response (CR) questions, and constructed-response questions with explanations (CRE, see Appendix¹). The CR questions require a short student response to the question. The CRE questions require a longer response, and students are not only expected to justify or explain the thinking behind their answers, but also to integrate personal knowledge and experience to extend the meaning. The MC and CR items on the reading component are scored on a 2-point scale (0, 2), and the CRE items are scored using item-specific scoring rubrics on a 3-point scale (2 marks for correct, 1 mark for partly correct, or 0 for incorrect). Reading abilities are defined by the EQAO in terms of reading with reasonable accuracy and proficiency in English: in other words, students are asked to connect relevant ideas and information so as to understand the meaning of the selected reading passages and to demonstrate moderate success in integrating their personal knowledge and experience to extend the meaning (EQAO, 2002).

EQAO reports of provincial results show that ESL/ELD students tend to fail the test and also to defer² writing the test at a far higher rate than the rest of the student population. For example, in October 2003 only 42% of the ESL/ELD students passed the whole test compared with an overall pass rate of 77%. About 45% of the ESL/ELD students passed reading only, and about 69% of the students passed writing only (compared with the overall rates of 82% and 88% respectively for all students who wrote the test). In terms of deferral rates, just over half of the ESL/ELD students (54%) participated in the test as compared with an overall participation rate of 91% (EQAO, 2003). Coupled with this higher deferral rate, the substantially higher failure rates of ESL/ELD students on the OSSLT suggest that this group of students is encountering great difficulty in meeting the graduation requirement.

Research acknowledges the potential effect of large-scale testing, noting that it has brought about both intended and unintended consequences to diverse groups of students (Madaus & Clarke, 2001). Minority students, including second-language students (such as ESL/ELD students in Ontario), are among the most vulnerable to the effects of such large-scale testing

policies (Shepard, 1991). These tests also tend to have more severe consequences for minority students and students from poor families (Horn, 2003; Madaus & Clarke). Two reasons potentially account for the more adverse effects of the OSSLT on ESL/ELD students than on non-ESL/ELD students. First, in terms of measuring English literacy development, ESL/ELD students may not be in a position equal to that of their non-ESL/ELD counterparts who have probably been part of the Canadian educational system for most if not all of their education and are likely as well to speak English as a first language. ESL/ELD students, however, may have been in the system only for a limited period before writing the OSSLT and are typically still struggling with the use of English as a second language. Researchers in second-language education suggest that four to eight years are required for ESL/ELD students to attain a level of language proficiency necessary to compete on a minimally competent level with their non-ESL/ELD counterparts (Collier, 1989; Cummins, 1981; Roessingh, 1999), and if ESL/ELD students have not had the time and experience to attain competent levels of English language-learning, they will be more likely to fail the test, with potentially negative consequences for their future academic studies or other pursuits. Another reason why ESL/ELD students may be more adversely affected by tests is because these tests were originally designed for non-ESL/ELD students, that is, students whose first language is English. Cornell (1995) has argued that evaluation standards that heavily rely on English-language skills are established for mainstream students; these standards overlook non-mainstream students' individual language progress (including second-language students), resulting in failure on tests governed by such criteria.

Literature Review of Test Format Effects

To attain validity and fairness in tests, efforts need to be made to minimize irrelevant effects on test performance (e.g., test format effects) and to examine if a given test measures the same construct across students with varied backgrounds (Bachman, 1990; Solano-Flores & Trumbull, 2003). Bachman emphasized the importance of research into test format effects on test performance, arguing that test developers could use information about interactions between test formats and test performance to help design tests that provide better and fairer measures of the language abilities that are of interest.

Various test formats have been argued to elicit varied levels of skill or ability. The multiple-choice (MC) format won its popularity in test design due to its scoring efficiency and freedom from ambiguity (Gay, 1980), along with its being "economically practical" and allowing "reliable, objective scoring" (Wainer & Thissen, 1993, p. 103). Nevertheless, many studies criticize the MC format. For example, the MC format has been challenged as inadequate to fully assess the dimensions of cognitive performance because

MC items provide limited opportunity to demonstrate in-depth knowledge (Fitzgerald, 1978); this format "may emphasize recall rather than generation of answers" (Wainer & Thissen, p. 103). In addition, there is the possibility that test-wiseness will contaminate the measurement. Test-wiseness includes a variety of general strategies related to efficient test taking (Bachman, 1990). With respect to the MC format, the strategy of ruling out as many alternatives as possible and then guessing among those remaining may be considered an example of test-wiseness.

The constructed-response (CR) format (including short-answer questions) is favored by some researchers and practitioners because it can measure traits that cannot be tapped by the MC format: for example, assessing dynamic cognitive processes (Bennett, Ward, Rock, & Lahart, 1990). Such items are also believed to replicate more faithfully the tasks test-takers face in actual academic and work settings. Furthermore, CR questions are considered to provide tasks that "may have more systemic validity" (Wainer & Thissen, 1993, p. 103). Because the CR format requires test-takers to construct their own answers, the assumption is that this format must involve higher-level thinking. But this idea has been challenged too. For example, Hancock (1994) investigated the comparative effectiveness of the MC and CR formats for assessing particular levels of complexity in the cognitive domain. He constructed examinations for two measurement classes with half MC and half CR questions. Equal numbers of questions in each format were written to reflect the first four levels of Bloom, Englehart, Furst, Hill, and Krathwohl's (1956) taxonomy.³ Hancock's argument was that given sound test construction, MC questions were able to measure the same abilities as CR questions across the first four levels of Bloom et al.'s taxonomy. The results indicated a pattern of highly disattenuated correlations between multiple-choice and constructed-response questions across increasing cognitive levels. Hancock inferred that ensuring that MC questions tap higher cognitive levels requires test constructors to have the necessary skills to develop distracters that reflect the desired cognitive level.

Based on the theoretical discussions above, earlier empirical studies have explored the influence of other test formats on students' performance. For example, Fitzgerald (1978) claimed that measurement procedures for evaluating the ability to comprehend written discourse (reading comprehension) were a critical concern in education. He investigated the differential performance of students at three grade levels in two cultures (United States and Irish) using three test formats: multiple-choice cloze, maze, and cloze. The results indicated that students from the two countries produced significantly different scores at grade 3 and grade 4. Concerning levels of difficulty for test formats, the study supported the assumption that MC questions would produce the highest student scores because they are recognition tasks. This result was upheld for both cultural groups. However,

cultural differences were also found. For example, the relatively higher scores on cloze items for Irish students were accounted for by an integrated program containing considerable creative writing. In contrast, the higher scores on MC cloze and maze items for US students were explained by their more skill-oriented and less integrated program. Thus it was concluded that the differential performance between the two cultural groups reflected characteristics of the educational programs in the cultures such as their different foci of orientation in developing reading skills or other sociolinguistic differences.

Following the thread of Kintsch and Yarbrough's (1982) study on the effects of test formats and text structure on reading comprehension, Kobayashi (2002) investigated the relationship between students' test performance and the two other variables: text types and test formats. She tested 754 college EFL (English as foreign language) students in Japan on four types of rhetorical organization: association, description, causation, and problem solution. Three test formats were employed: cloze, open-ended questions, and summary writing. Although the design of her study was challenged by some researchers (Chen, 2004), the results suggest that both text types and test formats had a significant effect on the EFL students' performance. Learners of different proficiency levels were differentially affected. Learners at higher English language ability were more susceptible to being influenced by different test formats. The results demonstrated that different test formats, including different types of questions in the same format, measured different aspects of reading comprehension. These findings also supported the concept of a "linguistic threshold" (Kobayashi, 2002, p. 210), according to which learners below a certain level of proficiency had difficulty understanding beyond sentence-level or literal understanding. Higher-proficiency learners, on the contrary, were more aware of overall text organization.

Shohamy (1984) investigated the effect of different testing methods, levels of reading proficiency, and languages of assessment on L2 reading comprehension by EFL readers. Using multiple-choice and open-ended questions presented in both the participant's first (L1) and second languages (L2), she found that learners performed better on multiple-choice questions presented in their L1, and these effects were greater for students with low levels of reading proficiency. Her findings indicated significant effects on students' scores in reading comprehension for all three variables: testing method, text, and language. In addition, Riley and Lee (1996) compared the summary and the recall protocol for reading performance by two levels of early-stage L2 readers of French. They asked half the participants to read a passage and then write a summary of the passage, and asked the other half to read the passage and then recall it. Findings indicated a significant qualitative difference in performance by the two levels of readers on the two tasks.

The major focus of earlier studies has been on the effects of test formats on student performance or assigned levels of proficiency. Based on these earlier studies, the present study aims to examine test format effects on students from different language backgrounds: ESL/ELD students and non-ESL/ELD students, that is, students who mostly use English as a second language and students who mostly use English as a first language. Given that ESL/ELD students are presumed to have a lower level of English language proficiency, the hypothesis in this study is that there would be a greater performance gap between ESL/ELD students and non-ESL/ELD students if they were required to integrate personal knowledge and experience to extend meaning in responses to CR and CRE questions. Two research questions guided this study. First, what are the performance patterns of ESL/ELD students on the three reading test formats compared with non-ESL/ELD students? Second, which test format(s) best distinguish(es) ESL/ELD students' performance from that of non-ESL/ELD students?

Methodology

Three sets of test data were obtained from the October 2003 administration of the OSSLT: (a) the 4,311 ESL/ELD students who wrote the test in October 2003, (b) 5,000 non-ESL/ELD students who passed the test, and (c) 5,000 non-ESL/ELD students who failed the test. From the non-ESL/ELD sample, a further random sample of students who either passed or failed the test were selected in conformity with the overall pass-fail ratio in the October 2003 administration (23% fail, 77% pass). To better represent the overall pattern, 77% of the non-ESL/ELD students (3,834 cases) who passed and 23% of the non-ESL/ELD students (1,169 cases) who failed were selected as a comparison with the ESL/ELD students ($n=4,311$). This resulted in a non-ESL/ELD student sample of 5,003.

Test scores on the reading component from the three formats—MC, CR, and CRE—were obtained from both student groups. There were 40 MC questions worth 80 marks, 35 CR questions worth 70 marks, and 25 CRE questions worth 50 marks. Descriptive statistics for the raw scores were first computed to determine the general patterns of ESL/ELD and non-ESL/ELD students' test performance. At the same time, other indicators—for example, standard deviation (SD), skewness, and kurtosis⁴—were obtained. Discriminant analyses were then performed to determine which format(s) could be used to distinguish the two groups. Discriminant analysis is most commonly used to classify cases into two or more groups based on various characteristics of cases and to predict group membership for new cases the group membership of which is undetermined (Norusis, 1988). The discriminant equation is $D = a + b_1X_1 + b_2X_2 + \dots + b_iX_i$, in which X_i represents each independent variable, b_i the corresponding coefficient estimated from the data, and D the predicted group membership. The resulting coefficients

provide the maximum separation among the groups. In this case the independent variables were MC scores, CR scores, and CRE scores, and D was predicted ESL/ELD membership. Subsequent correlation analyses were conducted to check if multicollinearity among the three test formats was a concern: that is, if MC scores, CR scores and CRE scores were highly correlated (.90 or above). Finally, classification results were obtained to demonstrate how well the discriminant functions differentiated the ESL/ELD students from non-ESL/ELD students.

Results

Descriptive analysis. The descriptive results show that both the ESL/ELD students and the non-ESL/ELD students obtained their highest mean scores in the MC questions (see Table 1): the correct percentage was 59.8% for ESL/ELD students and 74.1% for non-ESL/ELD students. Also, both groups obtained slightly lower scores on the CR questions: 58.2% for the ESL/ELD students and 72.7% for the non-ESL/ELD students. And both groups obtained their lowest correct percentage scores among the three formats on the CRE questions: 51.5% for ESL/ELD students and 65.2% for non-ESL/ELD students. The average differences on the three formats between the ESL/ELD students and the non-ESL/ELD students were 14.38% for the MC questions, 14.46% for the CR questions, and 13.68% for the CRE questions, indicating that the differences in performance were similar on the three test formats.

However, as indicated by the standard deviations (SD), the scores of the ESL/ELD students were more varied on the CR and CRE test formats than for non-ESL/ELD students. The standard deviations of ESL/ELD students on these two formats were 13.80 and 9.85, compared with 12.54 and 9.05 for non-ESL/ELD students. An examination of the skewness of the groups' scores demonstrated that the non-ESL/ELD students' performance was

Table 1
Descriptive Statistics of Reading Test Formats

	Mean (%)	SD	Skewness	Kurtosis
<i>ESL/ELD students</i>				
(n=4,311)				
MC (80)	47.87 (59.8%)	12.26	-.14	-.40
CR (70)	40.77 (58.2%)	13.80	-.49	-.34
CRE (50)	25.77 (51.5%)	9.85	-.33	-.48
<i>Non-ESL/ELD students</i>				
(n=5,003)				
MC	59.30/80 (74.1%)	12.46	-.84	.26
CR	50.89/70 (72.7%)	12.54	-1.11	.99
CRE	32.61/50 (65.2%)	9.05	-.86	.38

more negatively skewed regardless of format, indicating that non-ESL/ELD students' scores were more shifted to the higher end of the score distribution (higher scores). By examining the kurtosis value, it was found that the ESL/ELD students had negative kurtosis in all three formats (i.e., a flat distribution), indicating more spread in their scores, as opposed to the positive kurtosis obtained by the non-ESL/ELD students (i.e., a peaked distribution).

Discriminant analysis. Discriminant analysis was conducted to examine which test format had a better discriminating effect between ESL/ELD students and non-ESL/ELD students. The results show that only the MC test format was a significant predictor of group membership (see Table 2). The other two formats did not provide further significant separation between the two groups and are thus excluded from Table 2.

Large eigenvalues (relative proportion of variance contributed by each predictor) represent better discriminant functions. In other words, the ratio of the between-groups sum of squares to the within-groups sum of squares should be a maximum; in the current output the eigenvalue was .21, which was relatively small. The square of the canonical correlation (multiple correlations between predictors and groups) (0.42) and the difference in the value of Wilk's lambda from 1 (an index used to test the significance of the discriminant function) indicate that only 18% of the variance was associated with the differences between groups (see Table 2). Although the MC format provided significant distinction between ESL/ELD students and non-ESL/ELD students, a large proportion of the total variance was attributable to the differences within groups. In sum, the low eigenvalue coupled with the relatively high Wilk's Lambda indicated that although significant, the MC format did not strongly differentiate ESL/ELD and non-ESL/ELD students.

Follow-up correlational analyses revealed that the correlations among the three test formats—MC, CR, and CRE—were high (.84 between MC and CR; .80 between MC and CRE; .88 between CR and CRE). This explains why the other test formats were unable to discriminate group membership further in the presence of the MC results. Overall, the difference among test formats

Table 2
Discriminant Functions

	<i>Standardized Canonical Discriminant Function Coefficients</i>	<i>Wilk's Lambda</i>	<i>Eigenvalue</i>	<i>Canonical Correlation</i>	<i>F²</i>	<i>Sig.</i>
MC Score	1.00	.82	.21	.42	0.18	.001

Table 3
Classification Results

		<i>Predicted Group Membership</i>		<i>Total</i>
		<i>ESL/ELD</i>	<i>Non-ESL/ELD</i>	
Count	ESL/ELD	2,785	1,526	4,311
	Non-ESL/ELD	1,368	3,635	5,003
%	ESL/ELD	64.60	35.40	100.0
	Non-ESL/ELD	27.34	72.66	100.0

Note. 68.63% of original grouped cases correctly classified.

did not account for much of the variance between ESL/ELD students' and non-ESL/ELD students' OSSLT reading performance.

Given these results, it is not surprising that the classification results (see Table 3) also demonstrated that with the current discriminant function, test formats did not prove to be good discriminators in separating ESL/ELD students' and non-ESL/ELD students' reading performance on the OSSLT. Only 64.60% ESL/ELD students (2,785 out of 4,311) were correctly classified into their correct group, 14.60% above the chance level. Hence over 35% of ESL/ELD students (1,526) could be mistakenly grouped as non-ESL/ELD students. Similarly, only 72.66% of non-ESL/ELD students (3,635 out of 5,003) were correctly classified into their correct group, 22.66% above the chance level. Over 27% of non-ESL/ELD students (1,368) could be mistakenly classified as ESL/ELD students.

Given the discriminant function above, 68.63% of the original cases were correctly grouped (18.63% better than the chance level). Together these results suggest that test format is only a weak predictor of ESL/ELD membership.

Discussion and Conclusion

The results show that ESL/ELD students performed less well in all three test formats in the reading section than their non-ESL/ELD counterparts; however, the general patterns of difficulty were the same between the two comparison groups. Both groups achieved a higher percentage of correct answers in MC questions, lower in CR questions, and the lowest in CRE questions. These findings are partly supported by the literature; that is, MC questions are generally considered to be easier to answer correctly than CR questions or CRE questions (Fitzgerald, 1978; Shohamy, 1984). Students obtain higher achievement scores on MC questions than CR questions because MC requires "comprehension and selection," whereas CR requires "comprehension and production" (Wolf, 1993, p. 481). Furthermore, MC questions are

usually regarded as conducive to test-wiseness (Bachman, 1990). Such strategies may have resulted in the higher scores obtained on the MC format compared with the CR and CRE formats in this study. Further evidence to support this was provided by Cheng and Gao (2002), who found that in doing MC questions on reading comprehension, even in the absence of the associated reading passages, EFL students achieved scores above the chance level.

The finding that ESL/ELD students' performances were more varied than those of non-ESL/ELD students also has important implications. These ESL/ELD students, although all engaged in the English-language development process, vary considerably in their literacy achievement. Thus it is important that these results suggest that it may be necessary not to consider the ESL/ELD population as representing a homogeneous group, a common practice in school systems. In fact researchers and teachers may wish to pay more attention to examining individual differences among the ESL/ELD students instead of viewing them as single whole.

The discrimination analysis results combined with the unsatisfactory classification results based on the discrimination functions indicate that test formats provide weak discriminating power in separating the performance of ESL/ELD students and non-ESL/ELD students on the OSSLT. Given the current discrimination function based on test formats, approximately one third of the students could not be assigned to their correct group. Combining the results of descriptive analysis and discriminant analysis indicates that there are large performance gaps between ESL/ELD students and non-ESL/ELD students, but these gaps cannot be strongly attributed to test format differences. Cheng, Klinger, and Zheng (2007) conducted a two-year cross-validation study of the OSSLT data. Their results showed that the discrimination effect regarding test formats was not consistent over the two years of the study. For the February 2002 data, CR questions best separated the two groups $\beta=.42, p<.001$; MC questions had a discriminant coefficient of $.34 (p<.001)$, and CRE questions had the lowest discriminant coefficient of $.30 (p<.001)$. For the 2003 data (which are the same data as in this study), only MC questions separated the two groups $\beta=1, p<.001$. Also, Cheng et al. found that the performance differences were smaller in 2003 than in 2002. One possible explanation they offered was that the first test administration of the OSSLT had been in February 2002, whereas the October 2003 administration was the third. Thus the smaller performance differences could in part reflect the progress that ESL/ELD students had made or the extent to which these students had been coached for the test. Also, the smaller performance differences in 2003 might have led to the diminishing of the discrimination effects of the other two test formats, leaving only the MC format as a significant discriminator; therefore, test format did not provide a systematic separation between the ESL/ELD and non-ESL/ELD students. The dis-

criminating effects of test formats on test performance were significant yet weak, and the most difficult constructs (CRE questions) did not necessarily coincide with the best discriminator (MC questions) (Cheng et al.).

It is worth noting that the initial hypothesis guiding the current study was not supported in the findings: ESL/ELD students did not display noticeable extra performance discrepancies in CR or CRE questions compared with MC questions. Although a systematic analysis of the actual OSSLT questions is necessary to gain a deeper understanding of this result, possible explanations are offered. The CR and CRE questions on the OSSLT might not have required students to employ deeper cognitive levels, synthesis for example, or apply sophisticated background knowledge (which would have placed ESL/ELD students at a disadvantage) to answer the CR and CRE questions correctly. Thus further investigation would be justified with respect to whether the actual CR and CRE items on the OSSLT support the following two arguments from the literature reviewed above: (a) constructed-response formats are more advanced in assessing dynamic cognitive processes, as they are capable of asking students to employ not only knowledge-level but also synthesis-level cognition (Bloom et al., 1956); and (b) constructed-response formats replicate more faithfully the tasks that test-takers face in academic and work settings (Bennett et al., 1990). Subsequent studies could combine the analysis of test performance with further analysis of how test questions are constructed and answered by these students, that is, the reasoning and cognitive processes behind their choices and answers on the test.

Overall, the results of this study confirm that ESL/ELD students displayed substantial performance discrepancies compared with non-ESL/ELD students. These discrepancies, however, are close across test formats. The implication of this finding is that when teachers are preparing ESL/ELD students for the OSSLT, less focus should be put on the test format issue. Instead, with ESL/ELD students taking this large-scale provincial test while developing their English proficiency and literacy competence, a great deal of the variance in performance difference appears to relate to other aspects such as reading skills, reading strategies, or text types of reading passages, as indicated in Cheng et al.'s (2007) recent study. Thus ESL teachers' classroom priority should be their students' overall literacy competence rather than attention to test formats. This would include helping students to develop better reading skills and strategies and familiarizing them with reading the text types on the test (e.g., information, narrative, graphic).

Notes

¹In the sample OSSLT booklet, questions 1 and 2 are MC questions, question 3 is a CR question, and questions 4 and 5 are CRE questions.

²A deferral is made in consultation with the student and parents or the adult student, and with the appropriate teaching staff, on the basis that the student would not be able to participate in the test even with accommodations (EQAO, 2006).

³There are six cognitive levels in Bloom et al.'s (1956) taxonomy: knowledge, comprehension, application, analysis, synthesis, and evaluation.

⁴SD is a measure of variability; the larger the SD, the bigger the variance of the examined variable among the groups. Skewness and kurtosis are two measures usually reported to reflect the normality of the data. Normally distributed data have a value of zero for both kurtosis and skewness.

Acknowledgments

The authors acknowledge support from the Social Sciences and Humanities Research Council (SSHRC) of Canada and from the Education Quality and Accountability Office (EQAO) by releasing the October 2003 OSSLT data for this study.

The Authors

Ying Zheng is a doctoral candidate in the Faculty of Education, Queen's University. She has taught EFL in China. Her research interests are in ESL/EFL teaching and learning and large-scale language testing.

Liyang Cheng is an assistant professor in teaching English as a second/foreign language in the Faculty of Education, Queen's University. Her research interests focus on the effect of large-scale language testing on instruction and the relationship between classroom assessment and instruction in language classrooms.

Don A. Klinger is an assistant professor in assessment and evaluation in the Faculty of Education, Queen's University. His research interests include quantitative research methods, the examination of psychometric and policy issues of large-scale assessment, standard-setting, and measures of school effectiveness.

References

- Bachman, L.F. (1990). *Fundamental considerations in language testing*. New York: Oxford University Press.
- Bachman, L.F., & Palmer, A.S. (1982). The construct validation of some components of communicative proficiency. *TESOL Quarterly*, 16, 449-465.
- Bennett, R.E., Ward, W.C., Rock, D.A., & Lahart, C. (1990). *Toward a framework for constructed-response items*. ETS Research Report No. 90-7. Princeton, NJ: Educational Testing Service.
- Bloom, B.S., Englehart, M.B., Furst, E.J., Hill, W.H., & Krathwohl, D.R. (1956). *Taxonomy of educational objectives: The classification of educational goals. Handbook I: Cognitive domain*. New York: McKay.
- Chen, L. (2004). On text structure, language proficiency, and reading comprehension test format interactions: A reply to Kobayashi, 2002. *Language Testing*, 21(2), 228-234.
- Cheng, L., & Gao, L. (2002). Passage dependence in standardized reading comprehension: Exploring the College English Test. *Asian Journal of English Language Teaching*, 12, 161-178.
- Cheng, L., Klinger, D., & Zheng, Y. (2007). The challenges of the Ontario secondary school literacy test for second language students. *Language Testing*, 24(2), 185-202.
- Collier, V.P. (1989). How long? A synthesis of research on academic achievement in a second language. *TESOL Quarterly*, 23, 509-531.
- Cornell, C. (1995). Reducing failure of LEP students in the mainstream classroom and why it is important. *Journal of Educational Issues of Language Minority Students*, 15.
- Cummins, J. (1981). Age on arrival and immigrant second language learning in Canada: A reassessment. *Applied Linguistics*, 11(2), 132-149.

- Education Quality and Accountability Office. (2002). *Ontario Secondary School Literacy Test, October 2002: Report of Provincial Results*. Retrieved June 14, 2004, from: http://www.eqao.com/pdf_e/02/02P026e.pdf
- Education Quality and Accountability Office. (2003). *Ontario Secondary School Literacy Test, October 2002: Report of Provincial Results*. Retrieved June 14, 2004, from: http://www.eqao.com/pdf_e/03/03P006e.pdf
- Education Quality and Accountability Office. (2006). *Ontario Secondary School Literacy Test, 2006-2007: Guide for accommodations, special provisions, deferrals and exemptions*. Retrieved April 29, 2007, from: http://www.eqao.com/pdf_e/06/06P070e.pdf
- Elder, C. (1997). What does test bias have to do with fairness? *Language Testing*, 14(3) 261-277.
- Fitzgerald, T.P. (1978). A cross cultural study of three measures of comprehension at the primary and intermediate levels. *Educational Research Quarterly*, 3(2), 84-92.
- Gay, L.R. (1980). The comparative effects of multiple-choice versus short-answer tests on retention. *Journal of Educational Measurement*, 17(1), 45-50.
- Hancock, G.R. (1994) Cognitive complexity and the comparability of multiple-choice and constructed-response test formats. *Journal of Experimental Education*, 62(2), 143-157.
- Horn, C. (2003). High-stakes testing and students: Stopping or perpetuating a cycle of failure? *Theory into Practice*, 41(1), 30-41.
- Kintsch, W., & Yarbrough, J.C. (1982). Role of rhetorical structure in text comprehension. *Journal of Educational Psychology*, 74, 828-834.
- Kobayashi, M. (2002). Method effects on reading comprehension test performance: Text organization and response format. *Language Testing*, 19(2), 193-220.
- Kunnan, A.J. (1998). Approaches to validation in language assessment. In A.J. Kunnan (Ed.), *Validation in language assessment* (pp. 1-16). Mahwah, NJ: Erlbaum.
- Kunnan, A.J. (2004). Test fairness. In M. Milanovic & C. Weir (Eds.), *European language testing in a global context: Proceeding of the ALTE Barcelona conference, July 2001* (pp. 27-50). Cambridge, UK: Cambridge University Press.
- Madaus, G., & Clarke, M. (2001). The adverse impact of high-stakes testing on minority students: Evidence from one hundred years of test data. In M. Kornhaber & G. Orfield (Eds.), *Raising standards or raising barriers: Inequality and high-stakes testing in public education* (pp. 85-106). New York: Century Foundation Press.
- Ministry of Education and Training. (1999). *English as a second language and English literacy development: The Ontario curriculum grade 9 to 12*. Toronto, ON: Ministry of Education and Training, Canada.
- Norusis, M.J. (1988). *SPSS-X advanced statistics guide* (2nd ed.). Chicago, IL: SPSS.
- Roessingh, H. (1999). Adjunct support for high school ESL learners in mainstream English classes: Ensuring success. *TESL Canada Journal*, 17(1), 72-85.
- Riley, S., & Lee, J.F. (1996). A comparison of recall and summary protocols as measures of second language reading comprehension. *Language Testing*, 13(2), 173-190.
- Shephard, L. (1991). Negative policy for dealing with diversity: When does assessment and diagnosis turn into sorting and segregation? In E. Hiebert (Ed.), *Literacy for a diverse society: Perspectives, practices, and policies* (pp. 279-298). New York: Teachers College Press.
- Shohamy, E. (1997). Testing methods, testing consequences: Are they ethical? Are they fair? *Language Testing*, 14(3), 340-349.
- Shohamy, E. (1984). Does the testing method make a difference? The case of reading comprehension. *Language Testing*, 1(2), 147-170.
- Solano-Flores, G., & Trumbull, E. (2003). Examining language in context: The need for new research and practice paradigms in the testing of English-language learners. *Educational Researcher*, 32(2), 3-13.
- Wainer, H., & Thissen, D. (1993). Combining multiple-choice and constructed-response test scores: Toward a Marxist theory of test construction. *Applied Measurement in Education*, 6(2), 103-118.

Wolf, D.F. (1993). A comparison of assessment tasks used to measure FL reading comprehension. *Modern Language Journal*, 7(4), 473-489.

Appendix: Sample OSSLT Reading Booklet

Example Test Booklet Reading **5**

Get Rid of that T-shirt!


A recent newspaper article pointed out that Canadians purchased 73.7 million T-shirts last year. The article went on to say that the average North American owns 25 of them. The T-shirt was praised as the favourite garment of the twentieth century, worn by men and women, young and old, rich and poor. As we begin a new century, I suggest we leave the old T-shirt behind. 1

The first wearers of an "undershirt" or a "work shirt" in public were making a rebellious statement, but it quickly became the accepted style. Eventually, we all began to wear underwear anywhere and everywhere. 2

In the 60s, hippies tie-dyed their T-shirts. In the 70s, punk rockers shredded, safety-pinned and spray-painted them. In the 80s, T-shirts became great democratic portable billboards — each shirt an editorial column or personal ad telling others about the places the wearer has been, or the products, bands and politics the wearer supports or abhors. 3

The most recent trend seems to be toward slogans or messages that are increasingly meaningless. The best known examples are expensive T-shirts sporting only the name of the manufacturer. It seems strange that people are now expressing themselves by broadcasting their support of a shirt manufacturer. I can't think of anything less individualistic or less attractive to wear in public. 4

The T-shirt is basically a formless, ugly garment. What should happen to the 25 T-shirts each of us is supposed to have? I suggest that we use them as rags for washing our 1.7 cars. 5



multiple choice (Circle the letter next to the best or most correct answer for each question.)

1. In this selection, the T-shirt is compared to
 - A a slogan.
 - B a product.
 - C a garment.
 - D a billboard.

2. Which of the following is the best way to describe the purpose of this selection?
 - F to state an opinion
 - G to describe a product
 - H to present information
 - J to provide instructions

written answers

3. What is the meaning of the phrase "broadcasting their support of a shirt manufacturer" as used in paragraph 4?

4. Explain the purpose of the question in paragraph 5.

5. Do you think T-shirts will continue to be popular? Use one piece of information from this selection to support your answer.
