Walk a Mile in My Shoes: Stakeholder Accounts of Testing Experience with a Computer-Administered Test

Janna Fox & Liying Cheng

In keeping with the trend to elicit multiple stakeholder responses to operational tests as part of test validation, this exploratory mixed methods study examines test-taker accounts of an Internet-based (i.e., computer-administered) test in the high-stakes context of proficiency testing for university admission. In 2013, as language testing researchers (expert informants), we reported on our own experience taking the Test of English as a Foreign Language Internet-Based Test (TOEFL iBT) (DeLuca, Cheng, Fox, Doe, & Li, 2013). The present study extends these findings. Specifically, 375 current iBT test-takers, who had failed to achieve scores required for admission to university, completed a questionnaire on their test-taking experience. At the same time, two former test-takers who had passed the iBT volunteered for semistructured interviews. Questionnaire and interview responses were coded (Charmaz, 2007) for recurring and differentiating response patterns across these stakeholder groups. Concerns were shared regarding speededness, test anxiety, and test preparation, but these test-takers differed from the language-testing researchers in their responses to the computer-administered reading and writing tasks. Implications are discussed in relation to construct representation, the interpretive argument of the test (Kane, 2012), and test-takers' journeys through high-stakes testing to university study in Canada.

Conformément à la tendance de provoquer, auprès des parties prenantes, des réponses multiples aux tests opérationnels de sorte à valider ceux-ci, cette étude exploratoire à méthodes mixtes porte sur les récits de candidats à un test de compétence à enjeux élevés (l'entrée à l'université) et géré par ordinateur. En 2013, à titre de chercheurs en évaluation des compétences linguistiques (experts informateurs), nous avons fait état de notre expérience comme candidats au test d'anglais langue étrangère offert sur Internet (TOEFL iBT) (DeLuca, Cheng, Fox, Doe, & Li, 2013). La présente étude vient ajouter à ces résultats. Plus précisément, 375 candidats actuels au iBT n'ayant pas réussi à atteindre la note nécessaire pour être admis à l'université ont complété un questionnaire sur leur expérience lors du test. Deux autres candidats qui avaient réussi le test ont accepté de passer des entrevues semi-structurées. Les réponses au questionnaire et aux entrevues ont été codées (Charmaz, 2007) pour dépister des schémas récurrents et distinctifs parmi les groupes. Si les préoccupations relatives à l'effet des contraintes temporelles sur la performance, à l'anxiété et à la préparation avant le test étaient généralisées, les réponses aux tâches de lecture et d'écriture administrées par ordinateur n'étaient

pas les mêmes pour les candidats actuels que pour les chercheurs. Nous discutons des répercussions relatives à la représentation de concepts et à l'interprétation du test (Kane, 2012), d'une part, et aux parcours de candidats aux tests à enjeux élevés impliquant l'admission aux universités canadiennes, d'autre part.

> If we are to take seriously the argument … that the test-taker in particular and validation in general should be at the heart of development, then tests simply *must* be built around the test-taker. (O'Sullivan, 2012, p. 16)

With the global trend toward internationalization of university campuses and the cultural and linguistic diversity of Canadian classrooms (Fox, Cheng, & Zumbo, 2014), language tests have become ever more pervasive and more powerful decision-making tools (Shohamy, 2007). Inferences drawn about test-takers' language abilities based on language test scores result in life-changing decisions, for example, university admission, professional certification, immigration, and citizenship.

Across Canada each year, thousands of students enroll in English language programs and test preparation courses with the hope of improving their language and test-taking strategies in order to pass a high-stakes proficiency test. The present study took place in such a program, at a mid-sized Canadian university that enrolled students at basic, intermediate, and advanced levels during each 12-week term. Such programs have become a ubiquitous feature of the Canadian context (Fox et al., 2014).

Although test-takers are most directly affected by high-stakes proficiency testing, their role as the principal stakeholders in language testing has not always been recognized (Shohamy, 1984). In recent years, however, language testing validation studies have increasingly drawn on test-taker feedback in order to better understand how tests behave, and what they are actually measuring. For example, test performance has been researched in relation to test-taker accounts of

- test-taking strategies (Alderson, 1990; Cohen & Upton, 2007; Phakiti, 2008; Purpura, 1998);
- behaviours and perceptions before, during, and after a test (Doe & Fox, 2011; Fox & Cheng, 2007; Huhta, Kalaja, & Pitkänen-Huhta, 2006; Storey, 1997);
- prior knowledge (Fox, Pychyl, & Zumbo, 1997; Jennings, Fox, Graves, & Shohamy, 1999; Pritchard, 1990; Sasaki, 2000);
- test anxiety (Cassady & Johnson, 2002); and
- motivation (Cheng et al., 2014; Sundre & Kitsantas, 2004).

Studies have also drawn on test-taker accounts in order to examine what they reveal about a test method (Shohamy, 1984) or a task (Elder, Iwashita,

& McNamara, 2002; Fulcher, 1996). Others have elicited test-taker responses in consideration of a test itself (Bradshaw, 1990; Powers, Kim, Yu, Weng, & VanWinkle, 2009; Swain, Huang, Barkhoui, Brooks, & Lapkin, 2009).

Multiple-Stakeholder Accounts of Testing Experience

Recently, multiple stakeholder accounts of testing experience have been considered in the testing research literature as part of an ongoing program of test validation (Cheng, Andrews, & Yu, 2011; DeLuca, Cheng, Fox, Doe, & Li, 2013; Fox, 2003). For example, Fox (2003) examined differences in rater and test-taker accounts of a writing task in the context of the development and trial of a new version of a high-stakes test. Differences in these two stakeholder accounts led to a reconsideration of test specifications. Cheng et al. (2011) considered test-taking students' and their parents' perceptions of high-stakes assessment in Hong Kong. Qi (2007) compared students' and test developers' accounts of a writing test. Qi (2007) found differences in their perceptions of what was being measured, suggesting that understandings of a construct and what a test is actually measuring may differ in important ways from those intended by the test developers. Further, there is an ongoing need to accumulate validation evidence from operational tests in order to support the *chain or network of inferences* (e.g., Kane, 2012; Kane, Crooks, & Cohen, 1999; McNamara & Roever, 2006). This is what Kane et al. (1999) define as the "interpretive argument" (p. 6) of a test, that is, evidence that supports the interpretation or use of test scores. As Kane (2012) notes, validity itself is at its core an evaluation of the coherence and plausibility of evidence supporting a test's interpretive argument.

Messick (1996) pointed out that testing researchers and test developers should pay particular attention to construct underrepresentation and construct irrelevant variance as potential threats to the validity of inferences drawn from tests. In language testing research, however, once a test is operational, further consideration of these potential threats to validity has often been limited to an analysis of scores or outcomes alone (Bachman, 2000). Moss, Girard, and Haniford (2006) argue that validation studies should include stakeholder perspectives in order to expose sources of evidence that would otherwise stand to invalidate test inferences and uses. Bachman and Palmer (1996) also advise testing researchers and developers to explore test usefulness by eliciting feedback on operational versions of tests from key stakeholders (e.g., test-takers, raters, and other groups) who are affected by test decisions.

Over the past 10 years, we have seen an increase in the use of computers in large-scale language testing (see, for example, the TOEFL iBT or the Pearson Test of English [PTE] Academic). It is thus essential for testing researchers and developers to understand how the use of computers affects the test-taking experience and whether computer administration formats change the constructs being measured. Our review of the literature suggests that the role of computer-administered language tests in test performance is underresearched, in spite of their exponential growth. There has been some investigation of the impact of computer-administered testing (e.g., Maulin, 2004; Taylor, Jamieson, Eignor, & Kirsch, 1998), but it has arguably been insufficient. Fulcher (2003) noted the lack of published research on computer-administration interfaces in language testing in his consideration of a systematic interface design process. Since that time, some studies have been published that provide evidence to support the construct validity of various computer-administered tests in comparison with their paper-based counterparts (e.g., Chapelle, Chung, Hegelheimer, Pendar, & Xu, 2010; Choi, Kim, & Boo, 2003; Stricker, 2004), but these studies have tended to take place prior to the implementation of a new computer-administered test.

Computer-Administered Testing

A growing body of research suggests that we understand far too little of the implications of computer administration on the testing experience, of the ways in which a computer-administered format may subtly change a test construct (e.g., Hall, in press; Ockey, 2007), or the impact of the administration medium on test-taker perceptions and attitudes (e.g., Huff & Sireci, 2001; Richman-Hirsch, Olson-Buchanan, & Drasgow, 2000). Further, as Huff and Sireci (2001) note, "If the ability to interact successfully with a computer were necessary to do well on a test, but the test was not designed to measure computer facility, then computer proficiency would affect test performance." They go on to point out that "given that social class differences are associated with computer familiarity, this source of construct irrelevant variance is particularly troubling" (p. 19). Indeed, it is still the case that in a number of countries, students do not have wide accessibility to computers and are not accustomed to preparing their assignments using a computer.

As language testing researchers or expert informants (i.e., doctoral students and professors in language testing/assessment) who took the TOEFL iBT (hereafter, iBT; see DeLuca et al., 2013), we reported that our own experience with computer administration was generally very positive. For example, we recounted the ease with which we could respond to the writing task by typing our responses, and noted this as an improvement over the handwritten responses required of the paper-based tests we had previously written. Further, we reported that computer administration increased the overall sound quality of the listening and speaking sections of the test, and also allowed for the control of pacing in listening. At the same time, we identified "practical issues related to the language testing conditions, question design, and the testing protocol" (DeLuca et al., 2013, p. 663) of the test, which we argued were of potential concern with regard to construct. For example, we noted the high cognitive demands of the test (which we speculated were at times beyond those experienced by undergraduate students at the beginning of their degree programs). The cognitive demands were particularly evident in the reading section of the test (which was also the first section of the test).

We speculated that having such a difficult section at the beginning of the test might undermine the confidence of test-takers. Further, we expressed concern about the length of the test and the limited amount of time provided to complete complex tasks (i.e., speededness).

In order to extend and elaborate these findings, the present study elicited responses of former and current iBT test-takers—the target population/stakeholders of the test—and was guided by the following research questions:

1. What characterized the computer-administered testing experience for former (successful) and current (unsuccessful) iBT test-takers?
2. What did probing the testing experiences of these test-takers reveal about construct representation and the interpretive argument of the iBT? How do their accounts compare with those of the language testing researchers reported in 2013?

Method

The present study used an exploratory concurrent mixed methods research design (Creswell, 2015), merging or integrating findings from both qualitative and quantitative research strands. Notices of the study were posted in the university where the study took place in order to recruit former iBT test-takers who had recently passed the test (i.e., within two months) and were enrolled in their degree programs at the time of the study. Two students volunteered for semistructured interviews (see Appendix for the interview questions). At the same time, 375 recent iBT test-takers voluntarily responded to questionnaires circulated in 15 classes of a preuniversity English for Academic Purposes (EAP) program in the same university. None of these current EAP students/test-takers had passed the iBT at the time of the study. All of the questionnaire respondents had been required to upgrade their English proficiency in order to meet the minimum language proficiency requirements for admission to university degree programs.

Once the interview and questionnaire data had been analyzed, results from the two research strands were extended and explained by merging or integrating findings (Creswell, 2015)—a critical step in mixed methods research. In total, 377 test-taker participants contributed to the development of our understanding of what characterized these test-takers' testing experience with the computer-administered iBT, and how their experience differed or confirmed the accounts of the language-testing researchers reported in 2013.

Participants

Former test-takers ($n = 2$)

Two university students (former, successful iBT test-takers) were interviewed about their testing experience. Pseudonyms are used in reporting their accounts. Li spoke Mandarin as a first language (L1) and English as a second

(L2) or additional language. She had taken the TOEFL Paper-Based Test (PBT) in China, but her scores were not high enough to allow her to begin her Canadian university program. When she arrived in Canada, she completed an intensive, three-month iBT test preparation course prior to obtaining the TOEFL iBT test scores required for admission. Juan spoke Spanish (L1), English (L2), and French (L2). He had taken both the PBT and the iBT in the Dominican Republic prior to beginning his university program in Canada. Like Li, he had been unsuccessful on the PBT, but was later successful on the iBT.

Current test-takers (*n* = 375)

Current iBT test-takers voluntarily completed a questionnaire on their experiences taking the computer-administered test. Many of these participants indicated that they had taken a number of different proficiency tests, for example, the International English Language Testing System (IELTS) and the Canadian Academic English Language (CAEL) Assessment. All indicated that they were planning to take another proficiency test in the near future, but did not indicate which test they planned to take.

Instruments

The TOEFL iBT

Since its introduction in 2005, the iBT has been administered to millions of test-takers around the world. Although it is technically an Internet-based test, the present study focused on the computer-administered format or administration interface of the test (Fulcher, 2003). The iBT tests “academic English” in “reading, listening, speaking, and writing sections” (ETS, 2010, p. 6). It takes approximately 4½ to 5 hours to complete.

Interview questions

Semistructured interviews were conducted with the two former test-takers, who were asked to account for their testing experience (see Appendix for interview questions).

Questionnaire

The test-taker questionnaire used in the study combined items based on test preparation and the role of computers in test administration (DeLuca et al., 2013) and on the posttest questionnaire developed and validated by the Testing Unit at the university where the study took place. This questionnaire was routinely distributed after administration of high-stakes proficiency tests. The test-taker questionnaire was designed to elicit both closed and open-ended responses. Closed items collected information on key grouping variables. Open-ended items were the primary focus of the study. The questionnaire was controlled for length and complexity, given that it was administered across a wide range of language proficiency levels and only a limited amount of time was allowed by EAP teachers for administration of the questionnaire in class. The questionnaire is included in the Results and Discussion section below.

Data Collection and Analysis

As mentioned earlier, notices of the study were posted in the university where the study took place to recruit former iBT test-takers. Two students volunteered for semistructured interviews, which were audio recorded and transcribed for analysis. The questionnaire was circulated in the preuniversity EAP program at the beginning of a new 12-week term and across basic, intermediate, and advanced classes. Participants filled in the questionnaire in their EAP classes. Only participants who had taken the iBT within the previous two-month period were considered in the study. None of these participants had received scores high enough to allow them to enter their university programs at the time of the study. These current test-takers are the target population of the iBT. Further, all of these participants wrote a high-stakes paper-based test (i.e., the Canadian Academic English Language Assessment) under test conditions during the first week of the term. If their CAEL test scores had been high enough, they would have been deemed to have met the language proficiency requirement and admitted to their university programs.

The responses of the two former test-takers and the open-ended questionnaire responses of the current test-takers were analyzed using a modified constructivist grounded theory approach (Charmaz, 2006). Specifically, interviews were recorded and transcribed. Next, the texts were sorted and synthesized through coding, "by attaching labels to segments of data that depict[ed] what each segment is about" (Charmaz, 2006, p. 3). Through this process, the data were distilled by "studying the data, comparing them, and writing memos [to define] ideas that best fit and interpret[ed] the data as tentative analytic categories" (Charmaz, 2006, p. 3). Subsequently, the categories, which we identified in analysis of the interview data as *typified and recurrent features* (Paré & Smart, 1994) of the computer-administered testing experience, were compared with categories we identified in the coding analysis of open-ended responses to the questionnaire. In order to assess the reliability of the coding procedure, selected samples of interview and questionnaire responses were subsequently coded by two other researchers, who were familiar with the coding approach used in this study, but had not participated in it. Interrater/coder agreement was considered satisfactory based on Cronbach's alpha ($\alpha = .86$).

Quantitative data drawn from the questionnaires were analyzed using descriptive statistics (i.e., frequencies and percentages of response) to identify grouping variables. Open-ended responses were examined in relation to these variables (e.g., type of test preparation in relation to reports of test anxiety). Data from each of the research strands were analyzed and then integrated or merged in reporting the results. Merging the data from the two strands allowed us to extend and explain the findings with greater clarity and depth of interpretation. It should be noted that a distinctive characteristic and essential requirement of mixed methods studies (Creswell, 2015) is the integration of the separate findings from quantitative and qualitative strands. Given the present study's exploratory or naturalistic design, the qualitative findings are dominant, but they are more meaningful and interpretable when they are merged with the quantitative findings. Finally, we compared the language testing researchers' accounts of their iBT test-taking experience with those of the test-takers considered here, in relation to construct definition (Messick, 1996) and evidence supporting the interpretive argument of the test (Kane, 2012).

Results and Discussion

Overview

In exploring what characterized the computer-administered testing experience for both former (successful) and current (unsuccessful) iBT test-takers, we were also interested in any differences in their accounts, particularly those that might be considered construct irrelevant and potentially a threat to the interpretative argument for test use. We begin by presenting an overview of our findings, drawing on both the responses in semistructured interviews of the two former test-takers and the open-ended questionnaire responses of the 375 current test-takers.

Following Paré and Smart (1994), we reduced and synthesized the number of categories, identified as a result of multiple rounds of coding (Charmaz, 2006), into frequent and recurring themes that best characterize the computer-administered testing experience for the participants in the present study. The recurring themes across former and current test-takers are as follows.

- Acknowledgement of the importance of test preparation
- Concerns about speededness
- Positive responses to computer-administered tests of listening and speaking
- Mixed responses to reading subtests

In the section that follows below, we address our first research question in relation to these four recurrent themes:

1. *What characterized the computer-administered testing experience for former (successful) and current (unsuccessful) iBT test-takers?*

The Importance of Test Preparation

The questionnaire used in the study is presented in Figure 1, along with a summary of the responses (i.e., frequency and percentage, see square brackets) of the current test-takers to the closed items on the questionnaire. As

TEST-TAKER FEEDBACK QUESTIONNAIRE

Directions: Would you be willing to give us some feedback on taking English language tests like the TOEFL iBT, IELTS, or CAEL? If yes, please answer the following questions. Whether you answer or not, fold and drop this form in the box at the front of the room when you leave. Do not record your name. Thank you for your feedback.

1. Do you prepare in advance for an English language test like the TOEFL iBT, IELTS, or CAEL? **[*n* = 362 responses]**
   ( ) YES **[273, 75.4%]** ( ) NO **[89,** 24.6%]

   If YES, how do you prepare for a test? **Check all that apply**. **[*n* = 270 responses]**
   _**[200, 70.4%]** Look at the online practice tests
   _**[75, 27.8%]** Use the published Preparation Guide
   _**[79, 29.3%]** Take a preparation course
   _**[85, 31.5%]** Talk to friends
   _**[10, 3.7%]** All of the above
   Other: Please explain: ____

2. Have you ever taken an English test on computer? **(X) YES [375 or 100%]** ( ) NO

   If YES, check all that apply:
   __ TOEFL CBT
   _X_ TOEFL iBT [***N* = 375 or 100%]**
   __ Pearson Test of English (PTE) Academic
   __ Other (Please explain) ____

3. Why did you take the test(s) in #2 above?
   _X_ To get into university [***N* = 375 or 100%]**
   __ For my work
   __ For practice
   __ Other (Please explain) ____

4. Which method of testing do you prefer? **[*n* = 355 responses]**
   __ pen and paper **[302, 85%]**
   __ computer **[53, 15%]**

Please explain why you prefer this method of testing: ______________________

5. Do you think you would do better on the writing section of a test if you could use the computer to type your response? **[*n* = 315 responses]**

   ( ) YES **[92, 29%]** ( ) NO **[223, 71%]**

Please explain why: ____

Figure 1. Overview of current test-taker responses to the questionnaire.

indicated below, 273 (75.4%) of the 375 current test-takers who responded to the questionnaire indicated that they prepare in advance for a high-stakes test; 89 (24.8%) indicated they do not [13 (17%) did not respond].

In total, 270 (72%) explained how they had prepared for the test, with 117 (43%) indicating multiple approaches to test preparation, 200 (70.4%) mentioned accessing online resources, 85 (31.5%) indicated that they had consulted friends, 75 (27.8%) studied a test preparation guide, and 79 (29.3%) reported taking a test preparation course. Ten (3.7%) identified all of the above approaches as test preparation they engaged in prior to taking a high-stakes test.

Of the 117 (43%) respondents who indicated that they prepared in a number of different ways prior to taking a high-stakes test, the most frequently mentioned multiple forms of preparation were

- Look at the online practice tests and talk to friends: 28 (10.4%);
- Look at the online practice tests, use the published preparation guide, and talk to friends: 15 (5.6%); and
- Look at the online practice tests, use the published preparation guide, and take a preparation course: 12 (4.4%).

Because the current test-takers did not report their scores on the iBT, it was impossible to relate the amount or type of test preparation to a specific test or test performance. It was clear, however, that test preparation was an important feature in test performance for most of the current test-takers who responded to this item on the questionnaire.

One of the former test-takers also reported extensive test preparation prior to taking the iBT and explained how test preparation contributed to her use of specific strategies during the test. Li, the Mandarin-speaking test-taker, took the advice of a test-savvy friend and enrolled in "an expensive ($300) two-month, intensive iBT test preparation course" as soon as she arrived in Canada. She explained, "I recognized that I had to get to know the process and procedures of the iBT. I really needed lots to prepare. I had to get used to the computer." She added,

> I didn't use much the computer before in China, but when I come here [Canada] I realized I have to use computer for test. That actually created a lot of like high anxiety for me, because I don't know what is going on with the computer.

In addition to taking the course, she purchased a test preparation book describing the iBT and completed the book's activities and sample tests; used materials provided in her iBT registration package, both official (from the test developer's website), and unofficial materials, available online; and interacted with friends who had already taken the iBT. She also took workshops on using computers and practiced frequently in the library to improve her computer skills.

There is a well-reported tension (Cheng & Fox, 2008; Green, 2006) between language teachers' goals to improve their students' language in substantive ways, and students' goals to simply pass the test. Many students view such tests as barriers to their university education rather than as necessary verification that their language proficiency has reached the threshold essential for their academic work. There has been considerable concern that at times test preparation courses may potentially undermine a test's potential to measure language constructs of interest, and that test-takers may waste their time practicing test-taking strategies that are not useful beyond the bounds of test-taking itself (e.g., Cheng et al., 2011).

The comments of the test-takers in this study suggest that test preparation is indeed a large pretest focus in this high-stakes context. However, beyond the concerns the test-takers have for passing the test are the concerns relating to their development of computer skills that are adequate for the demands of the test itself. It is possible to make the argument that such skills are essential to academic work (and therefore part of the construct the iBT is measuring); however, one may question whether or not all entering university students have such skills at admission. This issue is discussed further below with regard to the iBT writing task, and in relation to the accounts of the language testing researchers reported in DeLuca et al. (2013).

Concerns About Speededness

Additional concerns about computer administration were evident in the responses of both former and current test-takers with regard to the amount of time provided to complete tasks on the iBT. Further to Huff and Sireci's (2001) misgivings regarding computer-administered tests, test-takers in the current study reported issues relating to time, the timing of tasks, and increasing test anxiety during the computer-administered test. Henning (1987) refers to this as *speededness*, a label we appropriated for this study. Similar results were reported in DeLuca et al. (2013) when, as test researchers/expert informants, we took the iBT and noted how demanding the test was. We reported feeling increasing anxiety and a sense of declining confidence as a result of the pressure to complete highly complex tasks within the imposed time limits even though there were no high stakes attached to our test-taking. The effects of speededness were frequently commented on in the present study as well. When current test-takers were asked to identify and explain their preferences for computer-administered or paper-administered tests, 302 (85%) indicated they preferred a paper-based format; 55 (15%) preferred the iBT's computer-administered format.

Amongst the 302 test-takers who preferred the paper-based format, time was explicitly mentioned by 31 test-takers in explaining why it was their preference (e.g., "takes more time on computer" or "more time to think and review with pen and paper"). Issues of time in relation to task completion were explicitly mentioned by 55 others, who explained their typing skills were "slow" or "poor"; 77 reported that paper-based administration was faster for them ("writing is faster with pen and paper" or "not as fast to write with computer"). Thus, issues of time and speededness figured in the preference for paper-based administration formats in 163 (44%) of the test-takers who responded to the questionnaire. Of the 55 who preferred computer administration, only 4 mentioned time as a reason; 7 explained they had good typing skills and 13 explicitly mentioned their typing was "faster."

In interviewing the two former test-takers, both of them mentioned that time and their speed of response was an issue for them in sections of the iBT. These comments are further explained below in relation to specific sections of the test.

Positive Responses to Computer-Administered Tests of Listening and Speaking

The benefits of allowing test-takers to control the overall pace of their work on the listening section of the iBT was frequently mentioned by both former and current test-takers (as was the case with the language testing researchers). Such control was not possible in paper-based formats. However, foremost in their positive responses to the listening and speaking sections of the computer-administered test were comments about sound quality and clarity, particularly in relation to the listening section of the test. For example, Juan (former test-taker) stated:

> I'm an ideal candidate for this study because I've taken the two versions of the TOEFL. I took both paper-based and iBT TOEFLs. So obviously I didn't get the score I needed when I took the paper-based test, but one of the reasons I had to take it again was the context in which it was administered. It was not fair. Specifically the listening section, because I was seated behind the speakers. It affected my concentration, because the quality of the audio was very bad and also the acoustics of the room.

He reported that his initial negative experience on the listening section of the PBT undermined his performance overall:

> I think my other scores on the PBT were lower because the listening was administered first. It just destroyed me. I didn't put as much effort after that. So listening failure triggered really high anxiety, and I was not motivated to write the other parts of the test. Well, I was motivated, but I couldn't concentrate. I was still thinking about the listening.

He received an overall score of 480 on the PBT and, as expected, his lowest score was in listening. Subsequently, he learned that the iBT was also being offered. He reported that when he took the iBT, "listening was my highest score" and "I was successful overall too, because I scored 105 on the iBT!"

He commented on the "quality of the sound, so easy, so clear" as a result of wearing headphones.

Improved sound quality and clarity in the listening section of the computer-administered test was explicitly highlighted by 7 of the 375 current test-takers as the clear advantage of the computer-administered test format. Like Juan, the 7 current test-takers who singled out sound quality as an issue for paper-based tests, reported test-taking experiences in which they perceived their performance had been undermined by conditions in the testing room itself: "I couldn't hear the sound of the lecture because I was in the back of the room" (Case 21) or "I couldn't hear the speakers" (Case 24).

However, not all of the issues related to sound have been resolved through computer administration. Many test-takers reported noise in the testing room as a problem in their test performance: "I couldn't think because the girl near me was on a different part [of the test] and she is speaking so loud. It is impossible to think and do my part" (Case 70). Other test-takers also complained about noise in the testing rooms: "the room too crowded and so noise is problem" (Case 26); "I can't hear because I hear other [test-takers] too" (Case 166). Several others reported distractions in the room: "In middle of test, this other one [test-taker] she has problem and makes noise and I can't focus on my test. Why they not take her outside to discuss?" (Case 12); or "new test-taker come into room to start test, but I not finished. I trouble then … can't think" (Case 10). The accounts of the current and former test-taker groups are very similar to those reported by the language testing researchers in 2013 with regard to ambient noise, further discussed below.

Mixed Responses to the Reading Section of the Computer-Administered Test

In this study, both the former and current test-taker groups reported that the reading section was not particularly difficult. For example, the former test-takers stated, "Reading was fine. No issues" (Juan) and "I liked the reading" (Li). They pointed out a computer feature that was helpful: "I could click on unfamiliar or technical terms. That helped me." And Li remarked that the multiple choice format made the reading section easier for her: "I guess because we're Chinese, we can use multiple choice. So you kind of have a strategy to exclude [distracters]. So that's my active [test] preparation strategy."

The comments of the former test-takers were similar to those of the current test-takers. It is important to note, however, that 39 (10.4%) of the 375 current test-takers reported that they "were not used to reading computer screens" (Case 27) under pressure; "feel nauseous when I read from computer" (Case 2); or "hate to read on computer [and] could not focus on the screen" (Case 33). Others reported they were "unfamiliar with reading on computer like this" (Case 200). Still others pointed out that they "like to underline and highlight," "like to circle keywords," and "write notes in the margin" when they read.

In order to explore this reported practice, and with permission of the Testing Unit at the university where the study took place, we examined 50 randomly selected reading booklets with extensive reading passages from previously administered CAEL Assessments. CAEL allows test-takers to work with and use the reading booklets while they are responding to questions on the reading subtest. Test-takers may write on the reading booklets if they choose to do so. The review suggested that when test-takers have reading booklets, a majority tend to annotate their reading in some way: 34 (68%) highlighted, wrote in the margins, circled, or underlined; 16 (32%) did not annotate the reading in any way. Further research on the academic reading construct needs to examine this finding.

Issues of construct representation are the focus of the section below, which addresses the second research question.

2. *What does probing the testing experiences of these test-takers reveal about construct representation and the interpretive argument of the iBT? How do their accounts compare with those of the language testing researchers reported in 2013?*

Probing the test-taking experiences of current and former test-takers (the target population of the iBT) suggests potential threats to validity that might not otherwise be evident if test-takers' perspectives are not consulted (Cheng et al., 2011; Fox, 2003; Fox & Cheng, 2007), or if test performances (scores) are the sole source of data, as has traditionally been the case with validation studies of operational tests (Moss et al., 2006). In the section below, construct-relevant issues identified by the test-takers' accounts of their computer-administered testing experience are discussed. Suggestions are made to further investigate these issues in order to determine how significant they might be, and to accumulate evidence with regard to the test's interpretive argument.

Responses to the Computer-Administered Writing Tasks

Ockey (2007) found subtle changes in construct as a result of a computer-administration format in the case of visual modes. In the current study, the changes were more dramatic, particularly in relation to test-takers' accounts of the iBT writing section. For example, the former test-takers identified the computer-administered writing section of the test as "the most difficult part" and potentially "unfair." Juan explained that he was "very anxious" about his "ability to type fast." He felt the computer-administered format of the writing section of the test put him at a particular disadvantage because "keyboarding was such an important requirement for writing well on the test." He explained:

> If I had been an expert in typing I would have performed better. This is the issue. If you are a slow typer [typist], the time is consumed by typing, and so you don't have time to go back and read and think about what you are writing. It steals, in a way, the time you would

have had to proofread and think about what you have written already. And if you look at it from that perspective, the ability to type fast on the keyboard, again definitely, I don't think that is fair in measuring whether you can write in English or not. Particularly for students who are from countries that do not have access to computers, or do, but do not type fast.

He pointed out that he drafts all of his university papers with pen and paper, typing only the final draft, because his "keyboarding is still very slow." He reported being "very nervous" during the writing section of the iBT, "because I type so slowly, I didn't have time to really finish or write what I wanted to say." He remarked that many of his current classmates and most students in his home country would not be able to perform well on the writing section of the computer-administered test. Like Li, they simply didn't have enough experience with computers or keyboarding. Nor did the former test-takers agree with the test researchers that familiarity with computers was ultimately an "advantage for studying in university." They pointed out that "familiarity" was different from "fast typing," and this skill was something that should not be expected of test-takers, who were just beginning university. They argued "it was unfair to expect this of second language students, when it is not expected of English-speaking students."

The former test-takers wondered why test-takers weren't given a choice to either type or write out their responses by hand. As Juan suggested, "Why don't they offer a choice for the writing section? Those who type quickly and write their papers this way could use the computer; those that don't, could use pen and paper, with the same amount of time to finish the work."

The responses of the current test-takers were also overwhelmingly negative about keyboarding requirements and/or the use of the computer for testing writing. In general, of the 355 test-takers who responded to a question about administration-format preference, only 53 (15%) indicated they preferred the computer, whereas 302 (85%) indicated they preferred paper-based administration for writing. Like Juan, they pointed out that "they were more used to pen and paper tests of writing" (Case 6). They argued that paper-based tests gave them "more freedom and mobility to write and erase manually … in computer you have to look at the keyboard" (Case 21).

Many of the current test-takers expressed concerns about controlling the computer, pointing out that "computer makes me nervous" (Case 28), and that a paper-based test is "safer than using computer. Sometime we press keys which can remove all we did" (Case 34). Still others expressed concern "because I am not fast in typing" (Case 23). They argued that "it's faster for a person to write on paper and reduces time" (Case 41) or "[it's] more natural" (Case 342), adding that "a lot of time [typing] means a lot of pressure on a person."

What stood out in our analysis of these test-takers' accounts were the differences in the amount and type of test preparation that they reported, because of the computer-administered writing tasks. Li's extensive preparation—particularly her extended emphasis on computer familiarity and keyboarding speed in preparation for the iBT—and Juan's comments on increased anxiety as a result of the demands imposed by keyboarding requirements of the writing task suggest that keyboarding speed and computer familiarity are part of the construct being measured by the test.

Juan (unlike Li) did not prepare for the iBT (having been unsuccessful on the TOEFL PBT and pinpointing listening as the problem, he registered without delay for the iBT). Although Juan did not do as well as he had expected on the iBT writing section, and during the test he experienced higher anxiety as a result of his lack of familiarity with typing his written work and his limited keyboarding skill, he still passed the test. Interestingly, within the unsuccessful (current) test-taker group, there were notable differences of opinion. Of the 315 who responded to the question, "Do you think you would do better on the writing section of a [high-stake] test if you could use the computer to type your response?" 223 (71%) responded "No" and 92 (29%) responded "Yes." In their explanations of their responses to this question, there were compelling differences between test-takers who reported their writing performance was undermined by typing their responses to the writing tasks, and those who reported that their performance was enhanced. This speaks again to the issue of construct, and what the test intends to measure. The accounts of the test-takers in the present study suggest a potential method effect, as the iBT requirement that they type their responses in the writing section of the test may differentially impact test performance.

In contrast, as language testing researchers or expert informants (i.e., doctoral students and/or professors in language testing and assessment) in the 2013 study, we reported that the iBT writing subtest was, in our view, the easiest section of the test. We appreciated the speed with which we could express our thoughts in writing as a result of the computer interface and complained about paper-based tests, which required us to write out responses, because we considered handwritten responses unnecessarily slow and limiting. Our positive accounts, as professional academics, of typing our responses on the writing subtest stand in sharp contrast to those of former test-takers, Li and Juan, and 75% of the current test-takers who are hoping to enter undergraduate programs. These differences may have important implications for construct definition and fairness.

This study suggests the need to further examine the impact of required keyboarding/typewritten responses on test performance. It should be pointed out that keyboarding is not a requirement for admission to English-medium universities in North America. Although the testing researchers in DeLuca et al. (2013) reported feeling at ease with typing their responses, by the time a student reaches graduate school (particularly at the doctoral level), typing written texts may be a "natural" part of academic work. This is not the case for entering first-year students. Most of the test-takers considered in the present study reported that the requirement to keyboard or type their writing impeded their performance on the iBT writing tasks. Their accounts of their testing experience raise construct (ir)relevant questions, given that typing original text under time pressure is not a requirement for admission to undergraduate universities in Canada. This is precisely the issue raised by Huff and Sirechi (2001), who note that students from contexts where computers are not a ubiquitous feature of education may be disadvantaged by this requirement. One may ask if it is fair to require this skill of only one group of entering undergraduates (L2 applicants), when others, who do not need to submit evidence of their language proficiency, do not face this requirement?

Ambient Distracting Noise in Testing Rooms

Although overall the test-takers considered in this study responded positively to the computer-administered listening and speaking sections of the iBT because of improved sound quality, ambient and distracting noise in testing rooms, as discussed above, was a frequently reported issue (also reported in DeLuca et al., 2013). Given that standardized test administration is foundational to measurement quality in large-scale high-stakes testing, this is an issue that would be of concern to test developers and test users, because it speaks to the interpretation of test results (i.e., the interpretative argument).

It appears that, in some test sites, administration logistics are well worked out to the advantage of test-takers. In other test sites, the close proximity of computer stations is a problem for some test-takers. Based on the varying reports of the test-takers considered here, there do not appear to be standard requirements for the positioning of computer stations/test-takers in a testing room (or, if standards are explicit, they may not be consistently followed). This needs to be systematically reviewed because it is a potential source of construct irrelevant variance. This finding could be investigated through the use of posttest questionnaires, which asked test-takers to comment on their testing experience. Over time, test-taker feedback would reveal administration issues arising in specific test sites that could then be addressed. In addition, requirements for test sites may need to be further detailed to ensure that logistics are comparable across test administration centres (e.g., that minimum distances between computer stations are respected, that activity in a test room is restricted, and so on).

Reading Extended Texts on Computer

Also of concern were the comments of current test-takers who reported that reading on a computer screen was either physically challenging (e.g., "made me feel nauseous") or unrepresentative of how they generally read academic texts (e.g., "I underline when I read," "I need to write notes when I read"). As Fulcher (2003), Huff and Sireci (2001), and Ockey (2007) have found, more research is needed to fully understand the impact and implications of such computer-administered tasks.

Although the construct of university-level academic reading was evident in the reported processes, procedures, and responses of the test researchers reported in DeLuca et al. (2013), it was not evident in the accounts of the former and current test-takers considered in the present study, who reported using test-wise (Cohen & Upton, 2007), multiple-choice test-taking strategies on the test—not academic reading strategies. Whereas the test researchers found the reading section the most demanding cognitively and commented on learning through their reading of the texts (albeit with concerns over insufficient time for reading in depth, i.e., speededness), most of the former and current test-takers seemed to take the reading subtest in stride—but not, it would seem, because they were effective academic readers. Rather, they reported using the multiple choice distracters strategically to find the "correct answers" (many of these practiced in test preparation courses prior to the test).

None of the former or current test-takers reported learning as an outcome of their testing experience, as had the testing researchers in DeLuca et al. (2013). The test-takers in the current study (the target group of the iBT) were reading strategically—for correct test answers, which one test-taker noted "were there, in the multiple choice options." This finding coincides with what Cohen and Upton (2007) found. Similar to Fox (2003), Cheng et al. (2011), and Qi (2007), these findings suggest that the construct intended by the test developer may not be the construct operationalized by the test, and may be undermining the interpretive argument for the test to a degree. This is important information for test developers, who may want to shorten the reading test (in keeping with the comments on speededness) and examine alternative response formats to avoid what appears to be a strong method effect as a result of the multiple-choice test format. In sum, whereas, based on the comments of the test researchers (DeLuca et al., 2013), an academic reading construct appears to have been operationalized by the reading section of the test, the comments of former and current iBT test-takers suggest that a different construct (unrelated to academic reading in university) may be operationalized for many in the iBT's target population.

Conclusion

This study investigated computer-administered testing experience by asking former and current test-takers for feedback on their testing experience. The results suggest that drawing on their insights increases our understanding of the operational test. However, the findings of this study must be interpreted with caution. First, the data for the study were drawn only from test-taker accounts of computer-administered testing experience. What we can account for and report is limited; so much of our experience is tacit. Further, our perceptions and accounts of an experience change over time, and the time between the test-takers' iBT testing experience and participation in the study was not fixed. Second, the questionnaire was administered only to unsuccessful iBT test-takers at the time of the study. All of these participants were volunteers. Their responses could not be linked to either iBT results or proficiency levels, which would likely have a bearing on the participants' views of the testing experience. Third, all of the data were collected from participants studying in one Canadian university, in either degree programs or in preuniversity EAP courses. Finally, only two of the former iBT test-takers who volunteered to be interviewed for the study met the criteria for selection (i.e., that they had not been successful on a high-stakes paper-based proficiency test, but had passed the iBT within the previous two months). If more former test-takers had been identified, they could have provided a much richer and thicker understanding of the computer-administered testing experience of successful test-takers. The interviews with the two former (successful) iBT test-takers were, however, clarified and extended by the questionnaire responses of the current (unsuccessful) test-taker participants in our study, and threw new light on the accounts of the iBT test taking experience reported by language testing researchers (expert informants) in 2013.

Despite acknowledged limitations, findings from this study suggest that the impact of computer administration on test performance needs to be further explored. More research is needed to address the threats to test performance and score interpretation posed by such issues as familiarity (test preparation), test method, speededness, and test anxiety, which we found in the current study, and were also raised by DeLuca et al. (2013), Huff and Sireci (2001), and Ockey (2007). Such issues speak to the interpretive argument of the test. As Kane, Crooks, and Cohen (1999) note, the ongoing collection of evidence drawn from *in vivo* or operational tests will either contribute to or lessen the meaningfulness of score interpretation and considerations of validity, which is essentially an evaluation of test interpretation and use. If, as suggested by O'Sullivan (2012) in the framing quote at the beginning of this article, the test-taker and validation are "at the heart of test development" (p. 16), then their accounts of testing experience are an essential source of test validation evidence. Test developers, test researchers, and other key stakeholders should also experience test-taking from the perspective of the test-taker. Walking a mile in test-takers' shoes provides important insights on how tests are measuring constructs of interest and their impact on test-taker performance.

The Authors

Janna Fox, PhD, Associate Professor in Applied Linguistics, Carleton University, teaches and undertakes research in language testing and curriculum, with a focus on diagnostic assessment and test validation. She received a 3M Teaching Fellowship for leadership in higher education and serves on the Board of Paragon Testing Inc., Vancouver, Canada.

Liying Cheng, PhD, is Professor and Director of the Assessment & Evaluation Group at the Faculty of Education, Queen's University. Her primary research interests are the impact of large-

scale testing on instruction, the relationships between assessment and instruction, and the academic and professional acculturation of international and new immigrant students, workers, and professionals to Canada.

References

Alderson, J. C. (1990). Testing Reading Comprehension Skills (Part Two): Getting students to talk about taking a reading test. *Reading in a Foreign Language, 7*, 465–504.

Bachman, L. F. (2000). Modern language testing at the turn of the century: Assuring that what we count counts. *Language Testing, 17(1)*, 1–42.

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

Bradshaw, J. (1990). Test-takers' reactions to a placement test. *Language Testing, 7*(1), 13–30.

Cassady, J. C., & Johnson, R. E. (2002). Cognitive test anxiety and academic performance. *Contemporary Educational Psychology, 27*, 270–295.

Chapelle, C. A., Chung, Y.-R., Hegelheimer, V., Pendar, N. & Xu, J. (2010). Towards a computer-delivered test of productive grammatical ability. *Language Testing, 27*(4), 443–469.

Charmaz, K. (2006). *Constructing grounded theory: A practical guide through qualitative analysis.* London, UK: Sage.

Cheng, L., Andrews, S., & Yu, Y. (2011). Impact and consequences of school-based assessment (SBA): Students' and parents' views of SBA in Hong Kong. *Language Testing, 28*(2), 221–249. doi:10.1177/0265532210384253

Cheng, L., & Fox, J. (2008). Towards a better understanding of academic acculturation: Second language students in Canadian universities. *Canadian Modern Language Review, 65*(2), 307–333.

Cheng, L., Klinger, D., Fox, J., Doe, C., Jin, Y., & Wu, J. (2014). Motivation and test anxiety in test performance across three testing contexts: The CAEL, CET and GEPT. *TESOL Quarterly*, *48*, 300–330. doi: 10.1002/tesq.105

Choi, I.-C., Kim, K., & Boo, J. (2003). Comparability of a paper-based language test and a computer-based language test. *Language Testing, 20*(3), 295–320.

Cohen, A., & Upton, T. (2007). "I want to go back to the text": Response strategies on the reading subset of the new TOEFL. *Language Testing, 24*, 209–250.

DeLuca, C., Cheng, L., Fox, J., Doe, C., & Li, M. (2013). Putting testing researchers to the test: An exploratory study on the TOEFL iBT. *System*, *41*, 663–676.

Doe, C., & Fox, J. (2011). Exploring the testing process: Three test-takers' observed and reported strategy use over time and testing contexts. *Canadian Modern Language Review*, 67(1), 29–53.

Elder, C., Iwashita, N., & McNamara, T. (2002). Estimating the difficulty of oral proficiency tasks: What does the test-taker have to offer? *Language Testing, 19*, 347–368.

ETS. (2010). *TOEFL iBT tips: How to prepare for the TOEFL iBT*. Retrieved from http://www.ets.org/Media/Tests/TOEFL/pdf/TOEFL_Tips.pdf. [Updated information now available at http://www.ets.org/toefl/ibt/prepare/

Fox, J. (2003). From products to process: An ecological approach to bias detection*. International Journal of Testing, 3*(1), 21–47.

Fox, J., & Cheng, L. (2007). Did we take the same test? Differing accounts of the Ontario Secondary School Literacy Test by first and second language test-takers. *Assessment in Education, 14*(1), 9–26.

Fox, J., Cheng, L., & Zumbo, B. (2014). Do they make a difference? The impact of English language programs on second language students in Canadian universities. *TESOL Quarterly*, *48*(1), 57–85. doi:10.1002/tesq.103

Fox, J., Pychyl, T. & Zumbo, B. (1997). An investigation of background knowledge in the assessment of language proficiency. In A. Huhta, V. Kohonen, L. Kurki-Suonio, & S. Luoma (Eds.), *Current developments and alternatives in language assessment* (pp. 367–383). Jyväskyla, Finland: University of Jyväskyla Press.

Fulcher, G. (1996). Testing tasks: Issues in task design and the group oral. *Language Testing, 13*(1), 23–51.
Fulcher, G. (2003). Interface design in computer-based language testing. *Language Testing, 20*(4), 384–408.
Green, A. B. (2006) Washback to the learner: Learner and teacher perspectives on IELTS preparation course expectations and outcomes. *Assessing Writing, 11*(2), 113–134.
Hall, C. (in press). Exploring computer-mediated second language oral proficiency testing: The test-taker's perspective.
Henning, G. (1987). *A guide to language testing*. Cambridge, MA: Newbury House.
Huff, K., & Sireci, S. (2001). Validity issues in computer based testing. *Educational Measurement, Issues and Practice, 20*(3), 16–25.
Huhta, A., Kalaja, P., & Pitkänen-Huhta, A. (2006). Discursive construction of a high-stakes test: The many faces of a test-taker. *Language Testing, 23*(3), 326–350.
Jennings, M., Fox, J., Graves, B., & Shohamy, E. (1999). The test-takers' choice: An investigation of the effect of topic on language test. *Language Testing, 16*(4), 426–456.
Kane, M. (2012, March). Validity, fairness, and testing. Paper presented at conference on conversations on validity around the world. Teachers College, NY.
Kane, M., Crooks, T., & Cohen, A. (1999). Validating measures of performance. *Educational Measurement: Issues and Practice,* 18(2), 5–17.
Maulin, S. (2004). Language testing using computers: Examining the effect of test-delivery medium on students' performance. *Internet Journal of e-Language Learning & Teaching, 1*(2), 1–14.
McNamara, T. & Roever, C. (2006). *Language testing: The social dimension*. Malden, MA: Blackwell.
Messick, S. (1996) Validity and washback in language testing. *Language Testing, 13*, 241–256.
Moss, P. A., Girard, B. J., & Haniford, L. C. (2006). Validity in educational assessment. *Review of Research in Education, 30*, 109–162.
Ockey, G. (2007). Construct implications of including still image or video in computer-based listening tests. *Language Testing, 24*(4), 517–537.
O'Sullivan, B. (2012). A brief history of language testing. In C. Coombe, P. Davidson, B. O'Sullivan, & S. Stoynoff (Eds.), *The Cambridge guide to second language assessment* (pp. 9–19). Cambridge, UK: Cambridge University Press.
Paré, A., & Smart, G. (1994). Observing genres in action: Towards a research methodology. In A. Freedman & P. Medway (Eds.), *Genre and the new rhetoric* (pp. 146–154). London, UK: Taylor & Francis.
Phakiti, A. (2008) Construct validation of Bachman and Palmer's (1996) strategic competence model over time in EFL reading tests. *Language Testing, 25,* 237–272.
Powers, D. E., Kim, H., Yu, F., Weng, V. Z., and VanWinkle, W. (2009). The TOEIC® speaking and writing tests: Relations to test-taker perceptions of proficiency in English*. ETS Policy and Research Reports, No. 78.* doi:10.1002/j.2333-8504.2009.tb02175.x
Pritchard, R. (1990). The effects of cultural schemata on reading processing strategies. *Reading Research Quarterly, 25*, 273–295.
Purpura, J. E. (1998). Investigating the effects of strategy use and second language test performance with high- and low-ability test-takers: A structural equation modeling approach. *Language Testing, 15*, 333–379.
Qi, L. (2007). Is testing an efficient agent for pedagogical change? Examining the intended washback of the writing task in a high-stakes English test in China. *Assessment in Education, 14*(1), 51–74.
Richman-Hirsch, W., Olson-Buchanan, J., & Drasgow, F. (2000). Examining the impact of administration medium on examinee perceptions and attitudes. *Journal of Applied Psychology, 85*(6), 880–887.
Sasaki, M. (2000) Effects of cultural schemata on students' test-taking processes for cloze tests: A multiple data source approach. *Language Testing, 17,* 8–114.

Shohamy, E. (1984). Does the Testing Method Make a Difference? The Case of Reading Comprehension. *Language Testing, 1,* 147–170.
Shohamy, E. (2007). Tests as power tools: Looking back, looking forward. In J. Fox et al. (Eds.), *Language testing reconsidered* (pp. 141–152). Ottawa, ON: University of Ottawa Press.
Storey, P. (1997). Examining the test-taking process: A cognitive perspective on the discourse cloze test. *Language Testing, 14*, 214–231.
Stricker, L. J. (2004). The performance of native speakers of English and ESL speakers on the Computer Based TOEFL and the GRE General Test. *Language Testing, 21*(2), 146–173**.**
Sundre, D. L., & Kitsantas, A. (2004). An exploration of the psychology of the examinee: Can examinee self-regulation and test-taking motivation predict consequential and non-consequential test performance? *Contemporary Educational Psychology, 29*, 6–26.
Swain, M., Huang, L., Barkhoui, K., Brooks, L., & Lapkin, S. (2009). *The speaking section of the TOEFL iBT™ (SSTiBT): Test-takers' reported strategic behaviors* (TOEFL iBT Research Report). Princeton, NJ: ETS.
Taylor, C., Jamieson, J., Eignor, D., and Kirsch, I. (1998). *The relationship between computer familiarity and performance on computer-based TOEFL test tasks* (TOEFL Research Report RR-98-08, TOEFL-RR-61). http://www.ets.org/research/policy_research_reports/publications/report/1998/hxwk

Appendix

Semistructured Interview Questions

1. I'd like to begin by asking you about your general experience with the test, your overall feeling, and your overall experience with this computer-administered test.
2. Were there any real stumbling blocks in your test-taking?
3. Perhaps I could ask you now about specific sections of the test.
4. Could you comment on your experience with the reading section of the test?
5. Could you comment on your experience with the listening section of the test?
6. Could you comment on your experience with the writing section of the test?
7. Could you comment on your experience with the speaking section of the test?
8. Which section(s) was the most difficult, and why?
9. Any final comments?