Available online at www.sciencedirect.com

ScienceDirect

System 41 (2013) 663–676

SYSTEM

www.elsevier.com/locate/system

Putting testing researchers to the test: An exploratory study on the TOEFL iBT

Christopher DeLuca [a,*], Liying Cheng [a], Janna Fox [b], Christine Doe [c], Miao Li [a]

[a] *Queen's University, Kingston, Canada*
[b] *Carleton University, Ottawa, Canada*
[c] *Mount Saint Vincent, Halifax, Canada*

Received 12 September 2012; revised 16 May 2013; accepted 18 July 2013
Available online 16 August 2013

Abstract

Despite being one of the most widely used proficiency measures of English for Academic Purposes, the newly-designed Internet-based Test of English as a Foreign Language (TOEFL iBT) remains externally under researched compared to previous TOEFL versions. Specifically, research is needed on factors that effect student performance within this new testing context. The purpose of this research was to identify and raise potential issues associated with the TOEFL iBT as explicitly linked to construct-dependent and construct-irrelevant variance factors. Through a key informant method, four language testing researchers sat for the TOEFL iBT and reported their experiences through two externally mediated focus groups. The testing researchers' experiences were analyzed using a standard thematic analysis then deductively grouped based on their form of measurement variance. Results provide valuable considerations for measuring English for academic purposes and serve to identify specific, practical issues related to the language testing conditions, question design, and the testing protocol of the TOEFL iBT.

Keywords: Language testing; TOEFL; English for academic purposes; Test-taking; Score variance

1. Introduction

At a recent academic conference, we, the authors of this paper, began discussing what it would be like to sit for the recently revised Internet-based Test of English as a Foreign Language (TOEFL iBT). As academics in the field of educational and language assessment, our discussion quickly turned to consider the linkages between language testing theory and practice, and the implications of theoretically informed measures for ensuring reliable and valid indicators of academic English language proficiency. Specifically, we wondered how the TOEFL's new Internet-mediated design, administration, and scoring would impact the test taking experience and contribute to score variance. In this context, score variance is understood as the difference between the students true score and his/her observed score on the test as attributed to a variety of testing factors.

* Corresponding author. A218 Duncan McArthur Hall, Faculty of Education, Queen's University, 511 Union Street, Kingston, Ontario, Canada K7L 3N6. Tel.: +1 613 533 6000x77675.
E-mail address: cdeluca@queensu.ca (C. DeLuca).

0346-251X/$ - see front matter © 2013 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.system.2013.07.010

In examining factors of test taking experiences and their relationship to score variance, Haladyna and Downing (2004) asserted the importance of collecting evidence that contributes to both construct-dependent score variance and construct-irrelevant score variance. Construct-dependent variance involves error in test performance attributed directly to measurement of the specific test construct (e.g., English for Academic Purposes), whereas, construct-irrelevant variance involves factors that are disconnected from the target test construct but that influence test performance. For example, the effects of an Internet-mediated test versus a paper-based test may skew scores depending upon the test-taker's abilities within these two testing contexts (Chapelle, 2008). Haladyna and Downing further recognize that systematic error arising from construct-irrelevant variance may result in unintended consequences directed to either all members of a particular examinee group (i.e., constant error) or differentially across test-takers. Accordingly, in studying test taking experiences and factors that may contribute to score variance, it is necessary to consider both construct-dependent and construct-irrelevant factors from multiple-perspectives, including those of the test-taker (Song and Cheng, 2006; Messick, 1998; Moss et al., 2006).

Building on this logic, we became interested in the experiences of test-takers as they sat for the TOEFL iBT and their perceptions of the construct-dependent and construct-irrelevant factors that may impact their test performance. Recently, there has been emerging research that explores test-takers' perspectives on test reliability and validity across language tests (Song and Cheng, 2006; Powers et al., 2009). However, few studies have been conducted on test-taker perspectives specifically on the TOEFL iBT. Given the relatively new and innovative Internet-based format and the widespread use of this test for high-stakes decisions (e.g., university admissions, scholarship, promotion), this perspective is critical to research in language testing and in better understanding the implications of this testing context on score variance.

In this study, we examine this perspective, but from a slightly different angle. Instead of soliciting the views of student test-takers, testing researchers (i.e., key informants) were selected to sit for the TOEFL iBT and report their experience through two post-test focus groups. We selected this method to mitigate some of the challenges traditionally faced in researching test-takers' perspectives; firstly, that test-takers do not have a strong understanding of testing theory and are limited in their articulation to accurately and explicitly connect their experiences to issues of reliability and validity, and secondly, that language test-takers often do not have a complete grasp on the intended measured construct (e.g., English for Academic Purposes) because the majority have not started university-level studies within an English-language context. This latter challenge may distort the accuracy of claims made in relation to construct-dependent factors. Accordingly, by using a key informant method (Patton, 2002), we felt that testing researchers would be able to more accurately articulate the linkages between testing experiences and language testing theory as related to construct-dependent and construct-irrelevant factors of score variance. In addition, having all completed advanced university studies within an English-language setting all participants were exceedingly familiar with the core construct measured by the TOEFL iBT, English for Academic Purposes (EAP). That said, we recognize at the onset of this study, that claims made about the TOEFL iBT are not intended to represent those of general TOEFL iBT test takers given the atypical and small sample size of participants in this study. Rather, our purpose in conducting this research is to identify and raise potential issues associated with the TOEFL iBT, as linked to testing theory concepts of construct-dependent and construct-irrelevant variance factors. In conducting this research, we intend to contribute to the ongoing dialog on language testing theory and practice, and specifically that of the TOEFL iBT, by offering yet another perspective to this growing area of research. This study is exploratory in design, aiming to lead toward a large-scale investigation. We assert this statement to mitigate attempts to generalize the study's results and further assert the caveat that this study is intended to provide initial insights into potential areas of TOEFL iBT development for more accurate measures of the EAP construct.

2. Background on the TOEFL iBT

Since its introduction in 2005, the TOEFL iBT has been administered to millions of test-takers around the world. Unlike its predecessors, the paper based (PBT) and computer-based (CBT) TOEFLs, it had as its goal to more effectively operationalize the construct of English for Academic Purposes (EAP) and increase alignment with the target domain of undergraduate university study. The test developer's intention for the iBT is (a) "to simulate university communication," (b) "to help test-takers determine their academic readiness," and (c) to help "institutions identify and select students with the English-communication skills required to succeed" in tertiary level study (ETS, 2010a,b, p. 5). Although multiple-choice test formats continue to dominate to differing degrees across the four tested

skills, there are extended speaking and writing tasks, both integrated (involving synthesis of or response to other texts) and independent, that are intended to facilitate more accurate measurement of the EAP construct (ETS, 2007, 2008).

The iBT begins with a reading section of, "3–5 passages, 12–14 questions each, 60–100 min" (ETS, 2010a,b, p. 6); followed by a 60–90 min listening section,"4–6 lectures, 6 questions each [and] 2–3 conversations, 5 questions each" (p.6). These two sections take approximately 2 h to complete, followed by a 10-minute break. After the break, test-takers return for the speaking and writing sections. Twenty minutes are allotted to 6 tasks on the speaking section: the two independent tasks provide 15 s of preparation time with 45 s to respond; the four integrated tasks have 30-second response times. Two of these tasks integrate reading and listening passages and allow 30 s for preparation; two relate to listening alone and allow 20 s for preparation. The 50-minute writing section is at the end of the test with both independent and integrated tasks.

The change from a traditional paper-and-pencil test to the computer-based format then to iBT has resulted in numerous innovations. For example, test-takers can now control the speed of the assessment and review previous questions in some sections of the test. The toolbar on the computer allows test-takers to control volume, provides a limited number of help features and a clock that informs test-takers of the remaining time in each section of the test. The test-taker can hide the clock if they choose at any time by clicking 'hide time' on the toolbar. Other innovations include the addition of visual items, increased flexibility in scheduling, more informative test reports, and faster marking and reporting of scores (Taylor et al., 1998).

The fact that the iBT is a high-stakes test for millions of test-takers annually provides the rationale for studies on the reliability and validity of the test (ETS, 2007, 2008). Although since 2000, various studies have been conducted on the construct validity of the test (Quinlan et al., 2009; Sawaki and Nissan, 2009), few studies have focused on test-takers (Cohen and Upton, 2006; Lawrence and Yigal, 2010), and yet, as the literature shows, test-taker accounts reveal qualities of a test that would not otherwise be evident, and that are critical to understanding factors associated with construct-dependent and construct-irrelevant variance (Fox, 2003).

3. Language testing and test-taker research

Despite the fact that studies on language testing have begun to consider the experiences and perspectives of test-takers (for a comprehensive review see Cohen, 2007), most have focused on test-taking strategies (Alderson, 1990; Cheng and DeLuca, 2011; Phakiti, 2008; Purpura, 1998), test-takers' behaviors and perceptions during test-taking processes (Elder et al., 2002; Lewkowicz, 2000; Storey, 1997), prior knowledge (Jennings et al., 1999; Pritchard, 1990; Sasaki, 2000), test-taking anxiety (Cassady and Johnson, 2002), and motivation (Sundre and Kitsantas, 2004). In general, these various studies have explored the factors that impact a test-taker's performance, rather than documenting and analyzing their perceptions related to testing experience and issues of test fairness, reliability, and validity (Song and Cheng, 2006).

In a discussion of validation in language testing, Messick (1996) asserted that test designers should collect a variety of evidence from various test stakeholders as related to test construct underrepresentation and construct irrelevant variance. More recently, Haladyna and Downing (2004) expanded Messick's validity framework with a classification of validation evidences related to construct-dependent and construct-irrelevant factors. Specifically, Haladyna and Downing (2004) argued that construct-irrelevant factors were largely understudied in large-scale assessment contexts despite their significant influence on issues of validity. In classifying specific aspects of test design, administration, scoring, and use, Haladyna and Downing noted that multiple factors impact test validity and that researchers need to build a more robust understanding of these factors. Moss et al. (2006) further argued that validation processes should include multiple stakeholder perspectives in order to expose sources of construct-dependent and construct-irrelevant evidence that would otherwise stand to invalidate test inferences and uses. However, in language testing research, consideration of these potential threats to validity has typically been approached solely from the perspective of the test designer (for a review see Bachman, 2000). There is a need to collect evidence from various test-takers about "how test-takers interpret test constructs and the interaction between these interpretations, test design, and accounts of classroom practice" (Fox and Cheng, 2007, p. 9).

Only recently have studies begun to address validity issues from the perspectives of test-takers, who, after all, are the principal stakeholders because they have the most at stake. For example, Powers et al. (2009) found evidence in using test-takers' perceptions to support interpretations of the new TOEIC® speaking and writing tests. They administered a self-assessment inventory to TOEIC examinees in Japan and Korea and gathered test-takers'

perceptions of their ability to perform a variety of everyday English-language tasks. Powers et al. (2009) found that TOEIC scores were consistent with test-takers' accounts of their performance of everyday language tasks in English, suggesting modest discriminant validity for the new measures of English-language proficiency. Another example is the study by Fox and Cheng (2007) that examined test-taker accounts of a high-stakes literacy test and found that the construct differed across L1 and L2 test-takers. This raised issues of construct fidelity (Loevinger, 1957) as the literacy processes engaged by the test differed in important ways from those in the target domain.

Cohen and Upton (2006) investigated the construct validity of the reading section of the TOEFL iBT based on 32 student participants. The reading tasks included the more traditional single-selection multiple-choice formats as well as the new selected-response (e.g., reading to learn items). They found that the reading tasks did indeed require participants to use academic reading skills, but the participants approached them as test-taking tasks rather than as academic tasks, because the desired task outcome was simply to get the answer right. The authors concluded that test-takers did not expect to learn or gain any new knowledge from the texts they read, and thus, the tasks were best considered as "test taking task[s] with academic-like aspects" (Cohen and Upton, 2006, p. 120). This study and the others cited here underscore the importance of test-taker accounts if we are to understand what our tests are measuring.

Bachman and Palmer (1996) also points out that although test developers routinely take their tests as part of the development process, once the test is operational (i.e., ready to be administered), they rarely elicit this important source of validation evidence. They thus advise testers to explore test usefulness by eliciting feedback from key stakeholders on operational versions of tests before and after tests are actually administered: "the more feedback obtained on usefulness, the more useful it will be" (p. 240). The present study explored what the iBT was measuring by eliciting accounts of four test-takers who were language testing insiders in a number of key ways of relevance to the iBT's construct. They provided feedback on the iBT's usefulness in response to the following research questions: How did testing researchers characterize their test experience? How did testing researchers describe the construct(s) measured through the TOEFL iBT? What, if any, potential sources of construct irrelevant variance or construct underrepresentation did testing researchers describe based on their test experience?

4. Method

This study used a key informant method (Patton, 2002) to ascertain the perspectives of testing researchers in sitting for the TOEFL iBT. Key informants are "people who are particularly knowledgeable about the inquiry setting and articulate about their knowledge—people whose insights can prove particularly useful in helping an observer understand what is happening and why" (Patton, 2002, p. 321). Linked to Eisner's (2004) evaluation concept of connoisseurship, key informants are able to discern nuances in experience and more accurately connect experience to theory, thereby enriching understanding of a phenomenon. Similarly, Stake and Schwandt (2006) noted, "as with connoisseurs and the best blue ribbon panels, some of the best examples of synthesizing values across diverse criteria are those that rely on the personal, practical judgment of fair and informed individuals" (p. 409). Accordingly, a key informant method was selected for this research to facilitate an articulation of the linkages between test-taking experiential factors and language testing theory as specifically related to construct-dependent and construct-irrelevant variance.

4.1. Participants

All of the participants in this study were researchers in language testing and educational assessment – two native speakers of English and two non-native English speakers of English. Three participants were in final stages of completing doctoral work in these areas and one was a professor in language assessment. We have used pseudonyms Mark and Mary for the two native speakers of English and Jiao and Dieu for the two non-native English speakers. Mark and Mary spoke English as their first language (L1) and were born and educated in Canada. Both Mark and Mary had conducted and published research in large-scale assessment, with Mary working for a testing company. Jiao, the other doctoral student participant, was born and educated in China prior to coming to Canada for her master's degree and subsequently PhD at an English-medium university. Jiao spoke Mandarin as her L1 and English as L2. She had taken the TOEFL PBT in 2005 in order to be admitted to her master's program, receiving a score of 620 at that time. The final participant, Dieu, was also born and educated initially in China, spoke Mandarin as L1 and English as L2. She had taken both an MA and PhD in English-medium universities, prior to working as a professor in Canada.

All four participants volunteered for the study because of their interest in the TOEFL iBT and because they met criteria for key informants related to this research topic: (a) researchers in educational assessment and language testing, (b) familiar with testing theory, and (c) personal experience with the target test construct (i.e., EAP) through graduate study education in English-medium universities. We recognize that participants of this study do not represent the typical TOEFL test-taking population; rather, they represent testing researchers external to ETS. Given their backgrounds and expertise in testing research we acknowledge that their perspectives, and the data collected in this study, are not reflective of typical TOEFL test-takers; they are perspectives as 'key informants,' which offers additional information about the test taking experience as articulated through an informed, exploratory, and theoretical perspective. In addition, we further acknowledge that the two native English speakers are linguistically different than typical TOEFL test-takers and as such, may have different responses to the test and that test items may function differently for these participants (ETS, 2010a,b). However, given their insight into the focal construct, we believe that these participants' English background provide important information on the test's ability to accurately represent and measure the EAP construct.

4.2. Data collection procedures

The four participants registered to take the TOEFL iBT at the same test center on the same day. They traveled to the testing center in another city and stayed overnight prior to taking the test because the test was not offered in their home city. Between registration for the test and the administration of the test, each prepared for the test on their own by consulting the TOEFL iBT website and other available test preparation materials. They downloaded free test preparation materials from the website, including a booklet with sample items and tasks. They also familiarized themselves, to the extent possible, with the test format and time requirements for each section of the test.

They arrived at the test center at 7:30 am and completed the test by 12:30 pm. Immediately after the test the participants met for the first of two focus groups. Both focus groups were moderated by the same moderator, who was another language testing developer and researcher. The focus group was semi-structured, audio-recorded, and lasted approximately 1.5 h. The focus groups were also observed by an external testing researcher in order to later review field notes and analysis from focus group transcripts. Questions for the initial focus group emphasized participants' experiences in sitting for the TOEFL iBT, including consideration of the test administration context, setting, and timing as well as test questions, presentation, and response formats. The second focus group took place two months later, once the 'test-takers' had received their scores on the TOEFL iBT. In this focus group, the moderator asked participants to comment on their scores and add insights they had about the test and their test-taking experience.

4.3. Data analysis

Data from the two focus groups were analyzed using a standard thematic analysis approach (Patton, 2002). First, data were analyzed in a 'zigzag process' (Creswell, 1998, p. 57) through open coding, in which sorting, coding, and comparison of data resulted in the identification of recurring themes in the responses of the participants in relation to the research questions. For example, during the first focus group, all of the participants recounted the challenges they faced in completing sections of the test within the time allowed. Recurring mention of such time constraints was labeled speededness and became a code. Codes were then clustered to form themes as the researchers re-examined and discussed the coded transcripts. Ultimately, speededness was positioned as a code within the theme labeled test design.

Themes and codes from the first focus group were then compared with those arising from the second focus group through a process of constant comparison (i.e., comparing data with data, interpretations, codes, categories and more data; Patton, 2002). Finally, in order to respond to the research questions and link participants testing experiences with language testing theory, themes were deductively grouped themes (Patton, 2002) as related to either construct-dependent factors or construct-irrelevant factors, following Haladyna and Downing's (2004) definition of these concepts and their classification of testing factors.

In order to triangulate the analysis, three approaches were used: (a) data were drawn from the participants on two occasions (i.e., analysis of data from the first focus group was verified through analysis of the second); (b) member checks, whereby participants responded to the findings, with their comments incorporated into the data; and, (c) an external testing developer and researcher, who conducts research and manages a high-stakes language test, observed the focus groups, took field notes, and discussed her comments on the data analysis with the moderator.

5. Findings

We present thematic findings in relation to construct-dependent and construct-irrelevant factors. Themes are elaborated upon using direct participant quotations with an emphasis on identifying potential issues associated with the TOEFL iBT, as linked to language testing theory. The quotations used in these results were representative of dominant trends in the data and are intended to provide examples of participants' accounts for articulating their testing experience. Table 1 provides an overall summary of the number of quotations coded for each theme.

5.1. Construct-dependent factors

Three construct-dependent themes were identified in participant data: (a) positive construct representation, (b) threats to construct representation, and (c) nontraditional construct dependent factors.

5.1.1. Positive construct representation

Evidence related to the theme of positive construct representation suggested that participants were all impressed with "the improved construct representation of academic language proficiency," as Mary noted with the agreement of the group, "the test and the passages looked very much like what we do at university." Indeed, there was consensus amongst the participants when Dieu interjected, "I really don't think construct is the issue. If you just look at the construct that has been embedded in the test, it's obvious they've done a lot in getting at this academic use of English." There was general agreement that the EAP construct conceived of by TOEFL designers was broad enough to encompass and match the use of English for university-purposes.

As recognized by participants, the test required the use of English in typical (and expected) ways, such as reading, writing, and listening, but also resembled alternative uses of English in academic contexts. For example, participants pointed out that they were required to use English to learn new concepts on the test. This learning process was perceived both as a benefit and as an authentic use of English for university purposes. During the first focus group, participants repeatedly engaged in lively exchanges about what they had learned. For example, Mary commented, "I learned something in that [reading] lecture." Dieu, agreeing with Mary, remarked, "Yeah I actually got the same thing. I now know the difference between *periods in the history of scientific thought*.[1] I really learned the differences through the test. That was amazing."

Mark pointed out that he noted an effective balance of topics in the test —"across arts, social science and science," adding, "for me, [the learning] sustained my interest a little more over those four and a half hours. If I was not able to learn, that would have seemed a much longer four and half hours!" Later, Mary again commented on her learning stated that she viewed it not only as evidence of appropriate construct representation but also of test fairness: "Going back to the learning you were talking about, if the item provokes learning it, it almost suggests that it's a…more fair item across [test-takers]. Because it means that we're at the same place as opposed to somebody coming in with background or prior knowledge and having that as an advantage to answering the question." Overall, the group felt very strongly that the learning potential on the test was evidence of construct representation and supported the claims of the test developer that the test operationalized the use of English on campus "from the classroom to the bookstore".

5.1.2. Threats to construct representation

Within this theme, participants identified several aspects of their testing experiences that were inauthentic or incomplete to the use of EAP. One widely discussed aspect was the use of paper-and-pencil note-taking sheets. While note-taking was certainly recognized by participants as construct-relevant, they also recognized that most students nowadays do not use paper-and-pencil to take notes during lectures but rather a computer program or a recording device. Further, Dieu reported that systematic note-taking during the test actually inhibited her performance: "I think the whole note-taking thing seems to be useless. That's what I realized, because I, I noted down a lot of details about the reading and listening, but none of those notes helped me to answer the specific questions!" In contributing to the conversation, Mary pointed out that although note-taking was not directly helpful in answering test questions, she noted that note-taking may have indirectly benefited her test performance: "I didn't actually go back and look at them

[1] The actual topics used on the iBT have been removed to insure that test content is not compromised. In order to preserve the meaningfulness of the comments, however, other topics have been inserted. These are italicized.

Table 1
Tally of quotations coded within each theme.

Relationship to construct	Theme	Number of quotations[a]
Construct-dependent	Positive construct representation	58
	Threats to construct representation	40
	Nontraditional construct dependent factors	28
Construct-irrelevant	Testing environment	39
	Test design: overall	95[b]
	Test design: reading	33
	Test design: speaking	23
	Test design: listening	25
	Test design: writing	29
	Score reports	56

[a] Represents number of quotations coded within a theme.
[b] The 'test design: overall' theme maintained a significantly higher number of quotations because specific test design quotations (i.e., reading, speaking, listening, and writing) were double coded.

[her notes] but may be writing out the things that I was hearing helped me remember them. So I mean they may have served a purpose, just not, in directly answering the questions." Mary's response supports previous ETS research on the value and correlation of note-taking with test-performance (Carrell, 2007). Carrell's study found that two note-taking approaches that consistently correlated with test performance were the number of content words recorded in the test takers' notes and the number of responses from the test recorded in the notes. Data from the present study confirm the first of these findings. In addition, our research relates test-based note-taking with the focal EAP construct. Dieu articulated that in university contexts, notes taken during lectures are typically augmented and/or synthesized with notes from readings and other sources to facilitate a more complete picture of the content with delineated and prioritized areas for studying purposes. Dieu noted that this complete process was not represented on the test, leading to an under-representation of the EAP construct and a potential misuse of note-taking strategies on the test.

Mark identified a second "inauthentic aspect of the test" when commenting on the reading section. He noted, "this isn't the way we read in university; we need to read carefully, reflect on what we are getting, and then oftentimes re-read." Adding to his comment, he stated, "if I'm reading an initial passage on *magpies* and I don't know words like *beak*, then I will go and look that up or I'll learn about that to get that background information, so then this passage does make sense to me and then I will respond to the questions and ask questions of my own, you know. But the test misses that sort of step. I think it's an important issue of construct underrepresentation." These two examples——note-taking and reading strategies——suggest that while the TOEFL iBT has touched on key aspects of the EAP construct, the current time constraints and format may limit a complete, authentic representation of the construct.

5.1.3. Nontraditional construct dependent factors

This theme relates to factors that have been identified in the literature as construct-irrelevant (see Haladyna and Downing, 2004) but that present an influence on the measurement of the intended construct for the TOEFL iBT. These factors included (a) the computer-mediated presentation and response format, (b) time sensitive responses, and (c) test length. Participants claimed that all of these factors reflected and shaped how English was used within an academic context, especially within university testing conditions.

In alignment with previous research (Wang et al., 2008), none of the participants found that taking the test on computer impeded their performance. However, these participants did identify linkages between computer use and the focal EAP construct: these participants viewed working on/with the computer as construct-relevant. As Mary noted, "we do this [work on computers] all the time." Computer-mediation as a facet of EAP is also noted in recent literature given that writing using a standard English-language (QWERTY)[2] computer keyboard is necessary to language use in English-medium universities (see, for example, Maulin, 2004). The computer screen and tools used in the TOEFL iBT were not standard, however, because they disallowed commonly used tools such as spell check, grammar check, or the

[2] This term refers to the standard arrangement of letters on keyboards (i.e., the first six letters, on the first row, far left, under the top row of numbers and symbols).

thesaurus, which would be used in EAP. Participants in this study made similar arguments for time sensitive responses and test length. While they noted that the test was taxing and that the timed responses were anxiety provoking, they recognized that they were typical protocols within university testing contexts and that constrained their quality and quantity of English usage. These latter two factors are further elaborated upon in the following section of findings.

5.2. Construct-irrelevant factors

In addition to identifying construct-dependent factors that were perceived to influence test performance, participants also identified notable construct-irrelevant factors that were aggregated into three themes: (a) testing environment, (b) test design, and (c) score reports.

5.2.1. Testing environment

In commenting on their overall test experience, all of the participants spoke of how "demanding" it was, "both physically and cognitively." Some of what made the test physically demanding may be traced directly to the amount of time and effort required to take the test. Like many test-takers, these participants needed to travel from another city, check into a hotel and stay overnight in order to be at the test center by 7:30 am. They completed their tests approximately 5 h later and all concurred with Mark who reported, "it was exhausting." The center where the participants took the test could accommodate up to six test-takers at a time in one small room, each in an individual cubicle with a computer and headphone.

Although none of the participants were particularly bothered by the surveillance in the room (e.g., security cameras trained on each test-taker, a window for administrators to monitor the room, mirrored ceiling, etc.), all reported being distracted by ambient noise and activity occurring in the room itself, as illustrated by Mark's observation: "what affected the test for me was that people were at different parts of the test at different times. So, when Mary was talking during the speaking portion, I could hear her." Other noise originated within the room as a result of administration procedures as Mary noted: "I missed part of my listening section because I was distracted by somebody who came in late for the test and they were talking." Other physical distractions and discomforts were reported including the need to wear headphones for 4 h, the lack of food and water throughout the test, and feeling cold due to having to leave shoes and coats outside the test area. Specifically, participants noted that they were not allowed to bring in any personal belongings: "we're not allowed to even bring in a tissue. I had a runny nose, and I was not allowed to bring a tissue" (Dieu), "we had to take our watches and shoes off and my feet were freezing for 5 h!" (Dieu), and finally Mary noted, "I felt very uncomfortable because they did not allow me to bring water, so after 1 h I felt very, very thirsty, but I had no choice. I had to still focus on the test."

5.2.2. Test design

In addition to articulating the effect of test environment of testing experience, the participants also spoke of the cognitive demands of the test as a result of the test's design. While participants noted specific design issues with each section of the TOEFL iBT (i.e., reading, listening, speaking, and writing), they further identified issues that affected all sections. Predominantly, participants noted the effect of *speededness* on their perceived test performance. Speededness refers to the length of the test or the timing of tasks within the test to respond adequately to complex questions (Henning, 1987). Mark commented, "I could not wait for the reading section to finish. It was very cognitively demanding. It was taxing." When asked if they experienced any test anxiety as a result of the timed responses, Dieu was the first to add: "I did have anxiety when I reached the second reading task and realized I had 12 or 13 questions left to answer within only 6 min. So I was very anxious." In response to this comment, all of the other participants nodded in agreement and each added their experience in feeling anxious and a sense of declining confidence as a result of pressure to complete highly complex tasks within the time limits of the test. For example, Mary stated, "that [timed responses] did give me that anxious jump of, 'oh my gosh, I'm not going to have enough time.' And, because when I first started the task I thought, well okay, I'm a competent native speaker, educated – I should do fairly well. It should be no problem for me. And then going through that first reading, and seeing that I only had so many minutes and so many questions left, I thought I'm not going to have enough time to finish them all!"

The participants also expressed concerns over the instructions on the test. They agreed that in some instances they were not only guessing at answers, but also guessing at the task requirements. Dieu stated, "They have this question where there are six sentences below a given text, but three sentences seem to be related and you needed to drag them

into the other column. But I didn't know if we had to drag them according to their sequence in the reading or if we just drag the relevant ones out of the six?" In response, Mark believed that sequence was unimportant, while Mary and Jiao thought that it was. Accordingly, the directions were unclear to these test-takers despite their initial test preparation, potentially creating construct-irrelevant measurement error and variance in test performance.

As the participants continued to engage with test questions, they began to identify test design patterns, which made directions seemingly more clear in some instances, but that led to confusion for other questions that did not follow the pattern. For example, in the ordering of questions (e.g., vocabulary items first, then main ideas) or the highlighting of key information (e.g. with all caps), these question formats were consistent in some sections but inconsistent in other sections. This mixed test design approach led Mark to attribute some of his confusion over task requirements to inconsistent signaling in instructions: "It [a key instruction] was in small writing, which was interesting because usually when they wanted to make a point they would use all caps. But in this instance I do not remember those instructions being in all caps. So it's confusing. Obviously it was a very important instruction."

Another salient issue raised by participants and directed most frequently at the reading section of the test, related to problematic distractors in the multiple-choice items. Repeatedly the participants described the multiple-choice questions as "tricky." They noted that at times, the distractors were "too strong." For instance, Mary commented, "those items that ask you to pick out the main points of the story, when they gave us six points that were in the story. But it's such a short passage, so the main points are not that clear. They could all be main points to some extent." The participants also complained that at times they found more than one correct answer or that distractors were ambiguous. Worse yet, they also reported that for some questions there appeared to be "no correct answer" (Dieu). In our discussion of distractor issues, we recognize that there may be relationship between item distractors and test construct, suggesting that distractor issues may be associated with construct-relevant factors. However, give participants' critiques that distractors were "tricky" and "ambiguous", we assert that the issue raised in these data is more accurately characterized as a construct-irrelevant factor because these critiques related to psychometric properties as opposed to 'construct'-dependent issues. That said, we do caution against generalizing this finding and encourage additional distractor analyses on the TOEFL iBT using psychometric and item review techniques. The data here serve to raise awareness of issues like this from test takers' perspectives. In addition to these various cross-cutting issues, participants identified specific design concerns related to each section of the test: (a) reading, (b) listening, (c) speaking, and (d) writing.

5.2.2.1. Reading. Although the participants agreed that the construct of academic reading was well served by the passages, they also felt that the difficulty level of the items might be beyond that encountered by first-year undergraduate students. Dieu pointed out, "it's much more complex reading compared with high school. Looking at high school students, I don't think they could handle it." Mary agreed and added, "it would be interesting to actually do a text analysis to see if the text was at a first-year undergraduate reading level." In addition, Jiao and Dieu raised issues of background knowledge and culture-specific knowledge that might undermine reading performance: "if you have some prior knowledge about one of the content areas then you can read it quickly. It's like you have more time. But, for example, the one about the *architect*. I'm not familiar with architects at all. So it took me a long time to read that passage, which meant I had only a short time to do the questions" (Jiao). Although Dieu found the two reading passages to be of a similar difficulty level, she supported Jiao's concerns about potential cultural bias in the *architect* passage: "if you grow up in China, you wouldn't know the *architect* described in the passage. You would actually have a lot of difficulty connecting the passage to art, as a foreign student." Mary acknowledged that what made some of the passages easy for her was "they were talking about things I knew." Although iBT/ETS test developers and researchers routinely inspect topics/tasks for potential bias, typically using a Differential Item Functioning analysis, it may be necessary to conduct additional studies across cultural groups, topics, and test methods to better understand issues of bias (see for example the work of Abbott, 2007, or Fox, 2003, for alternative approaches to bias detection).

5.2.2.2. Listening. Participants were generally positive about the listening section of the TOEFL iBT. They noted that test designers made a clear attempt to balance topics across science, humanities, and social sciences, and that test-takers could control the overall pace of their work. Although they could not stop the lecture segments, they could control the amount of time they took to answer the questions, giving them breathing space and reducing their anxiety. Further, Mary commented that the speed of delivery of the lectures was "normal, like a professor would speak."

Further, they all remarked that the lectures were carefully structured and well organized to support the listeners' comprehension. On the whole, the participants found the listening section fair and well designed.

Some concern was expressed with regard to the knowledge base required to understand a number of the lectures. Mary remarked: "what made it easy for me is they were talking about things I knew like, *Chaucer*. I understood because I studied English literature." Dieu commented that "for another foreign student, this may not be that straightforward," a point that was supported by Mark who added, "I learned that stuff during my undergraduate program, so if they are using this as an entrance to undergraduate, then that population of students wouldn't have as much background." As such, once again, participants suggested that the test content was positioned at an undergraduate level rather then a senior secondary school level, which is actually the intended population given that the TOEFL is an entrance test (Chapelle, 2008; Wang et al., 2008). While recognizing that some topics may be more familiar to some students than others, participants also acknowledged that little could be done about this from a test design perspective, as the test is taken by students at various levels from diverse educational backgrounds.

5.2.2.3. Speaking. None of the participants found the speaking section particularly difficult, but, as discussed above, *speededness* was a significant problem for all of them. This response suggests that the test's specifications regarding time on task may need to be re-evaluated. Not being able to respond quickly enough or to attain closure on the speaking tasks left these participants feeling "anxious" (Dieu), "dissatisfied" (Jiao), and "frustrated" (Mark), with the impact of this likely "greater for high-stakes test-takers" (Mary). Participants also commented that the integrated tasks (i.e., that synthesized information from readings and lectures) were generally easier, "and much more engaging" (Mark) than the independent tasks. Again however, time was an issue as participants reported that there was "not enough time to prepare for or respond to independent tasks" (Mary). They saw this as a gap in authenticity, as the independent tasks allowed 15 s to process what was being asked for and prepare a response, and only 45 s to articulate their response. They cautioned that given this experience, that there was a danger that speaking skills may be undervalued in the test design.

5.2.2.4. Writing. Similar to the speaking section, the writing section posed no serious challenges for the participants. However, the overall length of the test and the fact that the writing section was at the end may have influenced their perceptions, as Dieu pointed out: "I was feeling a little bit better because it's the last section. I was very happy to reach that section. So I think sequence of those tasks might make a difference to how we feel about the particular task."

The participants also generally agreed that the integrated task was more engaging and easier than the independent writing task. They commented that the independent task was too "simplistic", as Mark put it. "The integrated task was more reliant on some access skills in order for you to do well. But I found that the personal statement was, more challenging for me to write because the statement that they gave me didn't generate any sort of strong reaction in any way. I mean, 'do we feel it's better to spend money on a short holiday now or save for a long one later?' What do I care? They're both good, I'll take either" (Mark). Whereas in the other sections of the test, the participants would have liked more time, for the writing section they felt they could have completed the tasks with less: "if I had, I don't know, 5 min less, I probably could have still done the task" (Mary).

5.3. Score reports

Approximately eight weeks after the first focus group, the participants met to discuss their score reports. Dieu had correctly predicted that her lowest score would be in reading. In the first focus group, she had reported considerable difficulty with the reading task due to unclear instructions, mistiming, and some personal discomfort. Indeed, the score she received in the reading section of the test would have prevented her from being admitted to almost all Canadian university programs. This finding is concerning given her current position within an English-medium university and her previous educational background: She had already completed both MA and PhD degrees in English-medium universities and was engaged in a successful academic career. She speculated that she had probably missed all of the questions on the first task, pointing out that "preparation would definitely have helped me on the reading." Indeed, in addition to measuring reading ability, test preparation and construct-irrelevant factors (i.e., physical stamina, testing conditions, and test deign) may have contributed to Dieu's score.

As anticipated, the two L1 participants scored the highest on the test. However, Mark was surprised that he had not achieved a higher score in writing, which has "always been an academic strength for me." In analyzing this 'less than

perfect' outcome, he concluded that the result might be due to his response to the integrated task on the topic of *trout*: "there were two writing tasks and one of them was more of a summary. I think on the summary, I did a bang-up job. I should have, because I have a BSc degree and I studied *fauna*." Speculating on why he may not have done well on this task, he added, "I think I may have embellished the points with particular jargon that is subject specific, like the term *morphology*. I wrote a bit about morphology and its functional properties as well as other biological terms. I was conscious of who the raters might be, but for this task, that [word] wasn't in the text. So may be the rater wasn't capturing that——may not have known some of my terms. So that might have been perceived as *slight imprecision* (reading aloud from the criteria describing his performance on the score report)." He further articulated that the score report was fairly ambiguous and general, mitigating its utility for learning or other functional purposes.

Mary was surprised by her higher than expected score in reading because she recalled that "there was a lot of inference that wasn't readily apparent, some confusion and really complicated items. [It was] tricky." She was also surprised by her lower mark in listening, but on reflection added: "I do remember with the first listening task, that I had to guess on some of those questions and so perhaps that's what's reflected here." Mary also commented that it was difficult to know on the basis of the criterion-referenced descriptors how they might improve their performance. Like Mark, Mary remarked that the criteria were "simply too generic to be useful, and might even be misleading". Mark noted that he had received feedback that his writing might be "ungrammatical at times." He pointed out that when anyone writes under stress in a short period of time and without the benefits of standard editing tools, he/she inevitably "makes a few grammatical mistakes, but so what? Under those conditions, its understandable and expected." He speculated that such feedback on writing might mislead an L2 learner into needlessly focusing on grammar.

As she had predicted, Jiao's lowest score was in listening. "I hated the listening in this test," she said, comparing the iBT with the PBT, noting that the PBT allowed the test-taker more control and the potential "to guess answers even if you cannot understand." She also commented that her score on writing was lower than the mark she had previously achieved on TOEFL PBT writing. After two years of studying in Canada, completing a Master's degree, and a year of doctoral study in an English-medium university, "my writing has declined?" Like Dieu, Jiao's score may have been more impacted by construct-irrelevant factors, creating significant errors in measurement for Dieu.

6. Discussion and conclusions

This study put testing researchers to the test, in an aim to explore the TOEFL iBT test taking experience from a theoretically informed perspective. The four participants in the study were clearly not the target audience of the test, being well above the level of proficiency and academic expertise of the test's target population. However, as key informants (Patton, 2002), their expertise in both assessment theory and the focal construct, EAP, enabled a more articulate analysis of testing conditions, design, and measured constructs. Specifically, they were able to raise important issues that provide fertile ground for continued research on the TOEFL iBT or related computer-based language tests.

One of the most positive findings of this research was that participants agreed that the intended EAP construct was generally well represented by the TOEFL iBT. Of particular interest was that the test not only facilitated a measurement of typical EAP tasks (e.g., reading, writing, and listening) but also broadened the construct to include the utility of English for learning purposes (Chapelle et al., 2008; Educational Testing Service, 2007, 2008, 2010a,b). Although learning did not occur throughout the entire test, it was evident in some sections of the test and supports claims of construct authenticity. This finding is contrary to the research of Cohen and Upton (2006) and recognizes the importance of maintaining the complexity of the target construct within language tests. While we do not claim conclusively that a learning-orientation should be part of large-scale language tests or the measurement of the EAP construct, this finding does raise an important consideration for future research and test developers. Namely, we assert that it is necessary to continue to map out the EAP construct and find ways to assess the intricacies of the construct. For example, it may be useful to understand the extent to which computer-use and note-taking strategies are within the EAP construct and/or its sub-construct of EAP-listening, -reading, or -writing. Similarly, there is a need to examine whether or not time sensitive response or test length are perceived by test takers and test developers as potentially falling within the purview of EAP, as suggested by participants in this study. Engaging multiple-perspectives in this process, and by drawing on experts (such as was done in this research), will help inform the measurement of the EAP construct. A multiple-perspective approach will also help to address aspects of construct underrepresentation

(Messick, 1989, 1996) as identified by participants in this study (e.g., speaking section, note-taking approach, and reading strategy).

In addition to suggesting continued research on a larger scale and refinement of the EAP construct, this study contributes more generally to theoretical understandings of construct-dependent factors. Specifically, participants in this research noted that several factors that have been identified in previous research as construct-irrelevant (i.e., Haladyna and Downing, 2004; Kane, 2002; Messick, 1996) might in fact be construct-dependent factors for the EAP construct. For EAP, issues of timed-response, computer-mediated assessment, and test length constitute how English is used in university contexts, and in particular in university-level assessments. Accordingly, these factors are highly associated with English usage in academic settings. Thus, further to broadening the EAP construct, this finding suggests that original typologies of construct-dependent and construct-irrelevant factors must be interpreted in relation to the intended focal construct, and may be less rigid than originally postulated. This assertion may be most significant for the computer-mediated assessment factor with measures of academic readiness given the increased reliance on computers in all levels of teaching and learning.

Findings from this study also contribute to practical issues of the TOEFL iBT, which pose areas for future research and test development. These issues include:

1. *Level of texts*: Participants reported that many of the tasks seemed appropriate to the domain of university-level academic study, but that the level of some texts may be higher than that required for students entering university level of study. Additional research may be required to verify that text complexity is at an appropriate level, and that cultural bias is not playing a role in test performance. We recognize, however, that TOEFL iBT is also used for international students entering graduate programs.
2. *Timing*: While test length was perceived to be associated with the EAP construct, in the speaking and reading sections of the test, more time may be needed or the number of tasks reduced, in order to maintain the integrity of measuring speaking and reading skills. It is important for the test developer to consider funding research on test performance under different time conditions. The cumulative impact of being unable to complete tasks might undermine test-takers' overall test performance, leading to less valid results.
3. *Section order*. The participants found the reading section most difficult. It is also the first section that test-takers encounter. Reconsidering test choreography (Song and Cheng, 2006) so that easier tasks are at the beginning of the test might better scaffold test performance. The test developer may want to research the impact of re-ordering to see if it affects test performance.
4. *Testing environment and administration procedures*. When test-takers are experiencing discomfort and when they are distracted by ambient noise/activity, they may underperform on the test. The test developer may need to issue new guidelines on administration in order to insure minimum standards of comfort and to control ambient noise in order to mitigate this construct-irrelevant factor.

This study is one of only a few attempts to investigate large-scale language testing experiences from a key informant (i.e., assessment researcher) perspective. The findings of the study identified potential sources of construct-dependent and construct-irrelevant variance factors based on test-takers' 'live' test experiences. While results from this research have proven interesting and are useful to TOEFL test designers and to language testing researchers, they *must* not be interpreted as generalizable claims indicative of the typical TOEFL iBT test-taker population. We firmly assert that these results should be regarded as exploratory findings that identify potential areas for future larger scale research and test development. In particular, we see value in continuing to investigate the EAP construct with consideration for factors that would mitigate construct-dependent and construct-irrelevant variance. Ultimately, this research agenda would facilitate a greater basis for understanding test reliability and interpreting test score validity.

References

Abbott, M., 2007. A confirmatory approach to differential item functioning on an ESL reading assessment. Lang. Test. 24, 7–36.

Alderson, J.C., 1990. Testing reading comprehension skills (Part Two) getting students to talk about taking a reading test. Reading a Foreign Lang. 7, 465–504.

Bachman, L.F., 2000. Modern language testing at the turn of the century: assuring that what we count counts. Lang. Test. 17 (1), 1–42.

Bachman, L.F., Palmer, A., 1996. Language Testing in Practice. Oxford University Press, Oxford, UK.
Carrell, P., 2007. Notetaking Strategies and Their Relationship to Performance on Listening Comprehension and Communicative Assessment Tasks. ETS Research Report. Retrieved May 7, 2013 from: <http://www.ets.org/research/policy_research_reports/publications/report/2007/hsle>.
Cassady, J.C., Johnson, R.E., 2002. Cognitive test anxiety and academic performance. Contemp. Educ. Psychol. 27, 270–295.
Chapelle, C.A., 2008. The TOEFL validity argument. In: Chapelle, C.A., Enright, M.K., Jamieson, J. (Eds.), Building a Validity Argument for the Test of English as a Foreign Language. Routledge, London, pp. 319–352.
Chapelle, C.A., Enright, M.K., Jamieson, J. (Eds.), 2008. Building a Validity Argument for the Test of English as a Foreign Language. Routledge, London.
Cheng, L., DeLuca, C., 2011. Voices from test-takers: further evidence for test validation and test use. Educ. Assess 16 (2), 104–122.
Cohen, A.D., 2007. The coming of age for research on test-taking strategies. In: Fox, J., Weshe, M., Bayliss, D., Cheng, L., Turner, C., Doe, C. (Eds.), Language Testing Reconsidered. Ottawa University Press, Ottawa, ON.
Cohen, A.D., Upton, T.A., 2006. Strategies in Responding to New TOEFL Reading Tasks. TOEFL Monograph No. MS-33. ETS, Princeton, NJ.
Creswell, J., 1998. Qualitative Inquiry and Research Design: Choosing Among Five Traditions. Sage, Thousand Oaks, CA.
Eisner, E., 2004. The roots of connoisseurship and criticism: a personal journey. In: Alkin, M.C. (Ed.), Evaluation Roots: Tracing Theorists' Views and Influences. Sage, Thousand Oaks, CA, pp. 196–202.
Elder, C., Iwashita, N., McNamara, T., 2002. Estimating the difficulty of oral proficiency tasks: what does the test-taker have to offer? Lang. Test. 19, 347–368.
Educational Testing Service (ETS), 2007. TOEFL Research: Ensuring Test Quality. Retrieved August 13, 2008, from: <http://www.ets.org/Media/Research/pdf/Framework Recent TOEFL Research.pdf>.
Educational Testing Service (ETS), 2008. Reliability and Comparability of TOEFL iBT Scores. Retrieved October 2, 2008, from: <http://www.ets.org/Media/Tests/TOEFL/pdf/TOEFL_iBT_Reliability.pdf>.
Educational Testing Services (ETS), 2010a. TOEFL iBT Tips: How to Prepare for the TOEFL BT. Retrieved March 30, 2010 from: <http://www.ets.org/Media/Tests/TOEFL/pdf/TOEFL_Tips.pdf>.
Educational Testing Services (ETS), 2010b. TOEFL iBT Research Insight, Series 1, vol. 2. Retrieved May 7, 2013 from: <http://www.ets.org/s/toefl/pdf/toefl_ibt_insight_s1v2.pdf>.
Fox, J., 2003. From products to process: an ecological approach to bias detection. Int. J. Test 3 (1), 21–48.
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. Assess. Educ. Princ. Policy Practice 14 (1), 9–26.
Haladyna, T.M., Downing, S.M., 2004. Construct-irrelevant variance in high-stakes testing. Educ. Meas. Issues Pract. 23 (1), 17–27.
Henning, G., 1987. A Guide to Language Testing. Newbury House, Cambridge, MA.
Jennings, M., Fox, J., Graves, B., Shohamy, E., 1999. The test-takers'choice: an investigation of the effect of topic on language test performance. Lang. Test 16 (4), 426–456.
Kane, M.T., 2002. Validating high-stakes testing programs. Educ. Meas. Issues Pract. 21, 31–41.
Lawrence, S., Yigal, A., 2010. Test Takers' Attitudes about the TOEFL iBT. ETS Research Report. Retrieved May 7, 2013 from: <http://www.ets.org/research/policy_research_reports/publications/report/2010/ibmg>.
Lewkowicz, J.A., 2000. Authenticity in language testing: some outstanding questions. Lang. Test. 17, 43–64.
Loevinger, J., 1957. Objective tests as instruments of psychological theory. Psychol. Rep. 3, 635–694.
Maulin, S., 2004. Language testing using computers: examining the effect of test-delivery medium on students' performance. Internet J. e-Language Learn. Teach. 1 (2), 1–14.
Messick, S., 1989. Validity. In: Linn, R.L. (Ed.), Educational Measurement, third ed. Mcmillan, New York, pp. 13–103.
Messick, S., 1996. Validity and washback in language testing. Lang. Test. 13, 243–256.
Messick, S., 1998. Test validity: a matter of consequence. Soc. Indicators Res. 45, 35–44.
Moss, P.A., Girard, B.J., Haniford, L.C., 2006. Validity in educational assessment. Rev. Res. Educ. 30, 109–162.
Patton, M.Q., 2002. Qualitative Research and Evaluation Methods, third ed. Sage, Thousand Oaks, CA.
Powers, D.E., Kim, H., Yu, F., Weng, V.Z., 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. ETS, Princeton, NJ.
Phakiti, A., 2008. Construct validation of Bachman and Palmer's (1996) strategic competence model over time in EFL reading tests. Lang. Test. 25, 237–272.
Pritchard, R., 1990. The effects of cultural schemata on reading processing strategies. Reading Res. Q. 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. Lang. Test. 15, 333–379.
Quinlan, T., Higgins, D., Wolff, S., 2009. Evaluating the Construct Coverage of the e-rater® Scoring Engine. ETS Policy and Research Reports, No. 42. ETS, Princeton, NJ.
Sasaki, M., 2000. Effects of cultural schemata on students' test-taking processes for cloze tests: a multiple data source approach. Lang. Test. 17, 8–114.
Sawaki, Y., Nissan, S., 2009. Criterion-related Validity of the TOEFL® iBT Listening Section. ETS, Princeton, NJ. ETS Policy and Research Reports, No.81.
Song, X., Cheng, L., 2006. Language learner strategy use and test performance of Chinese learners of English. Lang. Assess. Quart. Int. J. 3 (3), 241–266.
Stake, R.E., Schwandt, T.A., 2006. On discerning quality in evaluation. In: Shaw, I.F., Greene, J.C., Mark, M. (Eds.), The Sage Handbook of Evaluation. Sage, Thousand Oaks, CA, pp. 404–418.
Storey, P., 1997. Examining the test-taking process: a cognitive perspective on the discourse cloze test. Lang. Test. 14, 214–231.

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? Contemp. Educ. Psychol. 29, 6–26.

Taylor, C., Jamieson, J., Eignor, D., Kirsch, I., 1998. The Relationship between Computer Familiarity and Performance on Computer-based TOEFL Test Tasks. TOEFL Research Rep. No. RR-61. ETS, Princeton, NJ.

Wang, L., Eignor, D., Enright, M.K., 2008. A final analysis. In: Chapelle, C.A., Enright, M.K., Jamieson, J.M. (Eds.), Building a Validity Argument for the Test of English as a Foreign Language. Routledge, New York.