State-of-the-Art Article

Recent developments in language assessment and the case of four large-scale tests of ESOL ability

Stephen Stoynoff Minnesota State University, Mankato, USA
stephen.stoynoff@mnsu.edu

This review article surveys recent developments and validation activities related to four large-scale tests of L2 English ability: the iBT TOEFL, the IELTS, the FCE, and the TOEIC. In addition to describing recent changes to these tests, the paper reports on validation activities that were conducted on the measures. The results of this research constitute some of the evidence available to support claims that these tests are suitable for their intended purposes. The discussion is organized on the basis of a framework that considers test purpose, selected test method characteristics, and important aspects of test usefulness.

1. Introduction

The purpose of this review is to provide readers with a survey of recent developments and validation activities related to four major international ESOL examinations: the Test of English as a Foreign Language™ (TOEFL®), the International English Language Testing System (IELTS), the First Certificate in English (FCE), and the Test of English for International Communication (TOEIC). More than half a million examinees take each of these tests annually, making them among the most widely used ESOL examinations in the world. The inferences and decisions made with the scores from these tests have significant consequences for examinees, score users, and society. Therefore, a review of these exams contributes to the professional discourse in several ways.

First, it provides a context for discussing some of the fundamental considerations and persistent issues in language assessment; these include (a) representations of the L2 construct, (b) validation theory, (c) applications of technology to language assessment, and (d) the consequences of assessment. Second, it reveals in concrete terms how language testing research and concerns for test consequences are affecting the test design and validation activities related to four influential assessments of L2 ability. Third, these tests have changed recently and this paper provides readers an update on each test. Finally, the following discussion is framed in terms of the test purpose, test method, and the evidence available to support the use of the test for its intended purpose. This approach reflects current practice and facilitates consideration of important aspects of the tests.

1 A list of acronyms used in the text will be found at the end of the article.
The review consists of an introduction, reviews of the four ESOL examinations, and a conclusion that suggests several issues worthy of additional research and future directions. Given the limitations of space, the paper focuses on the tests and even then does not attempt to include all the research available on them. Consequently, readers should consider the contents of this paper to be a personal perspective on the research related to these tests.

1.1 Representations of the L2 construct

Conceptualizations of the L2 construct influence the design and validation activities associated with large-scale tests. Since the early 1980s, L2 ability has been thought to consist of both linguistic and pragmatic knowledge. Canale & Swain’s (1980) work is widely considered pivotal in advancing conceptualizations of the L2 construct, but throughout the 1980s and 1990s Canale & Swain’s model was elaborated and extended by others, including Bachman (1990); Kramsch (1993); Bachman & Palmer (1996); and Chapelle, Grabe & Berns (1997). Presently, communicative L2 ability is thought to be comprised of a set of multiple subcompetencies that interact in a given language use situation. What is less clear and remains in dispute is exactly how many factors are involved and how they are related to each other (Chalhoub-Deville 2003).

The L2 construct has also been represented as a set of hierarchically arranged functions, situations, and the relative competence displayed by L2 users across a range of ability levels. The Common European Framework of Reference (CEFR) is an example of this orientation (Council of Europe 2001). It assumes that L2 ability occurs in a social context and is manifest through specific observable behaviors (language performance) in response to language tasks that are similar to what learners encounter in real-world contexts. This representation of the L2 construct has attracted much attention since its introduction and assessment specialists have debated the implications and limitations of this orientation (Fulcher 2004; Weir 2005a).

Another representation of the construct views L2 ability as performance of relevant tasks, and this perspective assumes that both linguistic and non-linguistic factors affect language use in a given context (McNamara 1996). Yet, the factors that affect performance on the task are complex and interact in different ways, and they are not fully understood. Some of the current concerns with performance-based task assessment are related to task difficulty, the adequacy of construct representation, and the ability to generalize from task performances (Bachman 2002; Norris 2002; Norris et al. 2002).

Finally, some researchers have suggested that the L2 construct ought to be defined as the interaction that language users engage in and the discourse they construct. This perspective considers language knowledge and performance as derived from the interaction of language, social contexts, and the co-construction of meaning. Test developers have encountered challenges in applying this perspective to assessment tasks, in part, because it is difficult to parse the contributions of participants to co-constructed discourse and to assign individual scores to it (Chalhoub-Deville & Deville 2005). Young (2000) contends that a theory of interactional competence can assist language assessment specialists in examining the particular configuration of interactional resources deployed by participants in a
given context of language use and offer ‘a principled way of generalizing from performance in the discursive practice of a performance assessment to performance in other non-test contexts’ (p. 11).

1.2 Validation theory

Validity is a crucial test quality and perspectives on it have evolved. Presently, validity is viewed from a unified perspective that includes an argument supported by multiple kinds of evidence that justify the score interpretations and use of the test for its intended purpose (AERA [American Educational Research Association], APA [American Psychological Association] & NCME [National Council on Measurement in Education] 1999). For many years, reliability was viewed as a discrete quality of a test; however, it is increasingly considered one kind of validity evidence and some reliability coefficients are used to support test developers’ claims of construct and content validity.

Much of the recent work in validation theory focuses on the development of an argument for test score interpretations and use, and the collection of different kinds of validity evidence. The work of Kane (1992, 2002), Mislevy and colleagues (Mislevy, Steinberg & Almond 2002; Mislevy et al. 2002), Weir (2005b), and others shaped the approaches to validation reported in the next section. Bachman (2005) acknowledges the contributions of argument-based approaches to the validation of score inferences, but contends considerations of test use and test consequences have not been incorporated into the development of validity arguments. To resolve this limitation, he proposes a more comprehensive approach to justifying test use, one that entails building an ’ASSESSMENT USE ARGUMENT’ consisting of two parts: ‘A UTILIZATION ARGUMENT, linking an interpretation to a decision, and a VALIDITY ARGUMENT, which links assessment performance to an interpretation’ (p. 16).

1.3 Applications of computer technology to large-scale assessment

In the past decade, the proliferation of computers and advances in computer technology offered new possibilities in test design, the analysis of responses, and the distribution of test results. Some of the most important developments with implications for large-scale assessments of L2 ability include additional (a) use of computer-adaptive testing; (b) attention to the contributions of INPUT and EXPECTED RESPONSE to task performance; (c) efforts to give examinees and test users more information about assessments and the scores they produce; and (d) use of technology in item development and the scoring of constructed response items (Jamieson 2005; Douglas & Hegelheimer 2007).

Despite the allure of technology and the surge of innovation, it is important to continue to raise fundamental questions about the application of technology to language assessment (Kern 2006). How do we make the best use of this technology in assessing L2 ability? How is technology-mediated communication similar to and different from other types of communication? The latter question has implications for developers of computer-delivered assessments as well as the examinees who take them. A study by Choi, Kim &
Boo (2003) used both content and statistical analyses to establish the comparability of examinees’ performances on computer-based and paper-based versions of an ESOL proficiency test. However, other studies are less conclusive. In a review of studies of examinees’ performance on computer- and paper-based versions of the same test, Leeson (2006) found that performance was not equivalent. The effects of examinee characteristics on performance (e.g., computer familiarity, computer anxiety, age, and gender) and test method characteristics on performance (e.g., item types, item presentation – font size or amount of white space, and interactivity – that is, the ability to scroll or review text) are not fully understood.

Computer-based assessment can enhance the content, scoring, and administration of tests, but it can also affect the validity of score inferences (Chapelle & Douglas 2006). In a review of recent technological innovations in large-scale computer-assisted assessment, Zenisky & Sireci (2002) noted advances in technology now make it possible to assess more skills, abilities, and processes than previously possible, but they cautioned that we have limited information on many aspects of this technology-mediated form of assessment. In particular, they indicated it was necessary to explore more fully (a) the complexity of a task and how it relates to the test construct being assessed, (b) the complexity or simplicity of the technology and how it affects the examinee’s performance, (c) the adequacy of construct representation and task generalizability, and (d) the extent to which automated scoring of constructed responses compromises score validity (Zenisky & Sireci 2002). Given the prominent use of technology in some high-stakes assessments, it is important for developers of these tests to demonstrate that the application of technology does not affect the generalizability of test scores.

1.4 Consequences of assessment

The term WASHBACK is considered by many to refer to the impact of a test on teaching and learning (Hamp-Lyons 2000). Recent investigations of the impact of high-stakes tests on teachers and classroom practices reveal that many factors are involved. Wall (2005) found this to be true in her study of the effects of introducing a high-stakes assessment into the educational system of Sri Lanka. Other recent studies indicate that it is not always possible to predict the effect of the test on teachers and language classrooms (Cheng, Watanabe & Curtis 2004; Qi 2005). Turner’s (2006) study provided additional support for this view. She surveyed ESOL teachers on what they perceived to be the effects of a high-stakes ESOL assessment on their classrooms and found the effect to be complicated and, in contrast with what is typically reported in the literature, that teachers considered the washback to be generally positive.

Major professional groups have responded to the concern for the consequences of assessment by codifying the standards and practices they advocate. The Code of fair testing practices (Joint Committee on Testing Practices 2004) is an example of a collaborative effort sponsored by the U.S. Joint Committee on Testing Practices (JCTP). Its members represent three principal professional associations, AERA, APA, and NCME, with the expertise to establish standards for educational assessments. The Association of Language Testers in Europe (ALTE) released its Code of ethics in 2000 and a draft Code of practice in 2001 (ALTE 2001). These publications specify the professional responsibilities of test developers and test
users, and if both parties assume their responsibilities, the codes will promote well-constructed tests and fair testing practices.

Kunnan (2004) posits that the notion of test fairness subsumes many of the consequences of test use. His five test fairness qualities consider validity, bias in test content and results, access, administration, and social consequences. Concerns for the potential bias in test content has raised questions about what norms to apply in international tests of English language ability (Elder & Davies 2006; Jenkins 2006a; Taylor 2006). The most widely used international assessments of ESOL ability assume examinees have acquired a variety of English that approximates the norms of Standard English. Are these tests biased against non-native speakers who have not been exposed to certain varieties of English? Davies, Hamp-Lyons & Kemp (2003) argue that there is a lack of empirical evidence to support claims of test bias with regard to major international ESOL exams, but others believe that exams such as TOEIC and IELTS privilege standard varieties of English and have ‘penalized examinees for using internationally-communicative forms of the language’ (Jenkins 2006b: 44). Canagarajah (2006) maintains that the debate over whose norms to apply to language use ought to be recast. Rather than trying to determine what constitutes a ‘universal proficiency’ he believes that the focus should be on developing assessments that measure context-specific language use in particular discourse communities (p. 235).

2. An examination of four large-scale, high-stakes tests of ESOL ability

Many different entities develop large-scale, high-stakes tests. However, as Eckes et al. (2005) found in the case of assessments of English language ability developed by states in the European Community, the standards and procedures applied to the design and use of these tests vary greatly. The following descriptions of the validation activities conducted by two major test development centers will highlight the current standards and practices applied to the development of high-stakes ESOL assessments.

A discussion of these tests can be organized in different ways. The approach taken here is based on an initial framework proposed by Read & Chapelle (2001) and elaborated in subsequent work by Chapelle, Jamieson & Hegelheimer (2003). It considers the test purpose, selected test method characteristics, and the evidence presented to justify use of the test for its intended purpose. The evidence is categorized in terms of the six qualities (reliability, construct validity, authenticity, interactiveness, impact, and practicality) of test usefulness proposed by Bachman & Palmer (1996).

In this approach, test purpose consists of the principal inferences to be made based on test scores, the decisions made on the basis of test takers’ performances on the test (or the uses of test scores), and the intended impact of the test on test users. The notion of test method is used here in much the same way that Bachman (1990) and Bachman & Palmer (1996) conceptualize it: a set of important characteristics of a language test, although only three aspects of their test method facets are considered in this discussion. They are the characteristics of the input, the characteristics of the expected response, and two aspects of the rubric – time allotment and method for response evaluation (for example, the number of correct responses or the level of performance as judged by a rater using a scale and scoring
criteria). Finally, test justification involves developing a validity argument that supports use of the test in a particular context, and it includes presenting compelling evidence that the test fulfills its intended purpose.

2.1 TOEFL

Since its introduction in 1964, the TOEFL (Test of English as a Foreign Language™) has evolved in response to the expressed concerns of score users, language testing specialists, and ESOL teachers as well as a result of the findings produced by a sustained program of research and development undertaken by Educational Testing Service (ETS) staff and outside experts. The first TOEFL was a five-part ‘OBJECTIVE’ multiple-choice test intended to assess ‘the English proficiency of non-native speakers of the language for academic placement in U.S. colleges and universities’ (Spolsky 1995: 3, 283). Based on the results of additional research, the five-section test was replaced in 1976 by a three-section test consisting of Listening, Structure and Written Expression, and Reading Comprehension. In 1986, a direct measure of writing ability (Test of Written English) was offered as a separate component that could be completed during some but not all administrations of the TOEFL. A computer-based version of the TOEFL was introduced in 1998. The Internet-based test (iBT) is the latest version of TOEFL, and it was launched in 2005 following a decade of validation activities intended to support the design and the proposed test score interpretations and uses of the test. According to a 2007 ETS news release (Educational Testing Service 2007a), approximately 750,000 examinees took the iBT in 2006 (http://www.ets.org).

2.1.1 Test purpose and method

According to The official guide to the new TOEFL iBT (Educational Testing Service 2006), the purpose of the test is to ‘measure English language proficiency’ (p. 1) and ‘the student’s ability to use English effectively in academic settings’ (p. 4). The publisher reports that TOEFL scores are used by over 5,000 entities in 90 countries to make a range of decisions apart from whether examinees possess sufficient English ability to succeed in North American colleges or universities (Educational Testing Service 2006). Nonetheless, given the nature of the language tasks, the input examinees receive during the test, and the focus of many of the research papers completed in conjunction with the development of the iBT, the primary target language use situation is North American colleges and universities.

The iBT has four parts (Reading, Listening, Speaking, and Writing), takes approximately 4 hours to complete, and is computer-delivered, although not computer-adaptive. There are some notable developments in the new test: the addition of a Speaking section, elimination of a grammar component, inclusion of integrated tasks (those that engage more than one language skill), adoption of new scales for reporting the Writing section score and total score, and availability of performance descriptions (competency descriptors) that can be used to interpret score information. ETS uses the Online Scoring Network (OSN), a secure Internet-supported system, to share data, train raters, and monitor raters’ scoring of examinees’
speaking and writing samples. Table 1 summarizes some of the key test method characteristics of the iBT.

2.1.2 Test justification

ETS spent more than a decade designing, investigating, and developing the validity argument for the iBT. As part of the validation process, the publisher commissioned and produced more than 40 monographs and technical papers between 1997 and 2007. Teams of applied linguists, educational measurement experts, and ETS staff researchers and test developers conducted the research. These publications constitute a valuable source of expert knowledge on many of the most pertinent issues in large-scale L2 testing and they are available on the ETS TOEFL website. Additionally, a case study of the iBT project was published recently and it offers readers a comprehensive description of the validation activities conducted during the development of the validity argument for the new TOEFL (Chapelle, Enright & Jamieson 2008).

2.1.2.1 Construct validity

Previous versions of the TOEFL did not include a theoretical rationale to support the interpretation and uses of test scores. One of the early papers completed for the iBT project produced a working purpose statement and a model of the overarching construct developers sought to assess; it was referred to as the Committee of Examiners model or COE (Chapelle et al. 1997). A subsequent paper articulated a general plan for undertaking important elements of the project (Jamieson et al. 2000). Four additional papers extended and translated the general test domain articulated in Jamieson et al. (2000) by further specifying the listening (Bejar et al. 2000), reading (Enright et al. 2000), speaking (Butler et al. 2000), and writing (Cumming et al. 2000) constructs for the various sections and features of the new test. In addition to contributing to the theoretical rationale to support the interpretation and use of test scores, these papers informed some of the empirical investigations that provided evidence for the TOEFL validity argument.

In a review of the Listening section of the TOEFL Computer-based Test (CBT), Buck (2001) noted the listening construct was underrepresented, in part, because the oral input to examinees was too inauthentic and the tasks did not have examinees apply what they heard to a target language use situation. The iBT Listening section, based on the COE model and work completed for Monograph 19 (Bejar et al. 2000), includes more and longer samples of spoken discourse and new item types (complex tasks) that assess important aspects of the listening construct. Current theoretical conceptualizations of the reading construct informed the content and characteristics of the tasks used in the iBT Reading section (Enright et al. 2000). In an analysis of the CBT Writing section, Weigle (2002) concluded that the writing construct was somewhat narrowly defined because the test task was limited to demonstrating the ‘ability to write argumentative discourse on an impromptu topic’, and it did not include writing for different purposes, audiences, or genres (p. 146). The writing construct is defined
<table>
<thead>
<tr>
<th>Component</th>
<th>Input</th>
<th>Response</th>
<th>Time</th>
<th>Evaluation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reading</td>
<td>3–5 passages of about 700 words each</td>
<td>Select single correct response from 4 choices</td>
<td>60–100 minutes</td>
<td>Number of correct answers (1–4 points each)</td>
</tr>
<tr>
<td></td>
<td>12–14 questions per passage</td>
<td>Select all correct choices from set of 3–7 possibilities</td>
<td></td>
<td>Reported on scale of 0–30</td>
</tr>
<tr>
<td></td>
<td>Some questions include statements and table</td>
<td>Select correct choices and complete table or summary</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Listening</td>
<td>4–6 (3–5-minute) mini-lectures heard once (may contain one or more visuals)</td>
<td>Select 1 (or 2) correct response(s) from 4 choices</td>
<td>60–90 minutes</td>
<td>Number of correct answers (1–2 points each)</td>
</tr>
<tr>
<td></td>
<td>6 questions with choices (written input)</td>
<td>Select correct choices and complete chart</td>
<td></td>
<td>Reported on scale of 0–30</td>
</tr>
<tr>
<td></td>
<td>2–3 (3–5-minute) conversations heard once</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>5 questions with choices (written input)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Speaking</td>
<td>6 tasks</td>
<td></td>
<td>20 minutes</td>
<td>Level of performance as judged by trained raters (3) using scale (0–4) and assessment criteria (delivery, language use, and topic development)</td>
</tr>
<tr>
<td></td>
<td>2 independent tasks: spoken and written instructions/prompt</td>
<td>15 seconds to prepare and 45 seconds to respond</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>4 integrated tasks</td>
<td>20–30 seconds to prepare and 60 seconds to respond</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Task Type</td>
<td>Task Description</td>
<td>Time</td>
<td>Reporting</td>
<td></td>
</tr>
<tr>
<td>----------------------------------</td>
<td>----------------------------------------------------------------------------------</td>
<td>--------</td>
<td>-----------</td>
<td></td>
</tr>
<tr>
<td>Type 1 task: read/listen/speak</td>
<td>Respond to questions related to the situation or academic topic</td>
<td></td>
<td>0–30</td>
<td></td>
</tr>
<tr>
<td>Spoken instructions and prompt</td>
<td>75–100-word reading passage followed by 90 seconds of related discourse</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Type 2 task: listen/speak</td>
<td>Summarize or demonstrate understanding by responding appropriately</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>90 seconds of discourse</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Writing</strong></td>
<td>2 tasks</td>
<td>50 minutes</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Independent task: written prompt</td>
<td>Compose 300-word essay expressing an opinion on the topic</td>
<td>(20 minutes)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Integrated task: written prompt</td>
<td>Compose 150–225-word summary based on the input</td>
<td>(30 minutes)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Level of performance as judged by trained raters (2) using scale (0–5) and assessment criteria (content, development, language use, and organization) or on scale of 0–30.
more broadly in the iBT, and it includes pragmatic considerations, additional rhetorical functions, multiple writing samples, and complex tasks that require the use of several language skills (Cumming et al. 2000).

iBT developers recognized the limitations in using the computer to deliver the speaking component of the new test (Butler et al. 2000). Natural oral interaction involves direct and immediate exchanges with others and participants in the interaction typically assume the roles of speaker and listener in a dynamic process that results in jointly constructed discourse. However, current technology is not capable of replicating this experience in a computer-delivered task. After carefully considering the practical implications of administering a more authentic speaking assessment, developers designed a semi-direct assessment to elicit speaking performance. The speaking construct assessed in the iBT was defined in terms of the knowledge and control of important phonological and syntactic features, vocabulary, and discourse patterns encountered in academic contexts.

While it was important to establish a theoretical basis for both the design and the proposed interpretations and use of iBT test scores, developers also sought to collect empirical evidence to support the validity argument. One aspect of this work investigated the kinds of listening, reading, speaking, and writing tasks considered important to succeed in university studies. As part of these domain analyses, Rosenfeld and his colleagues asked a sample of practitioner informants drawn from 22 North American universities to judge the relative importance of task statements related to the four language skills and academic performance in undergraduate and graduate courses (Rosenfeld, Leung & Oltman 2001). The results represented some of the evidence used to justify the design of the iBT and to support the claim that test content was relevant and representative of what examinees encounter in university contexts. Researchers also collected and examined authentic spoken and written academic language use (Biber et al. 2002). Biber and his colleagues’ corpus analyses of spoken data informed the design and content of the input used to assess listening comprehension, and their work provided support for the claim that iBT test content approximates what is encountered in the target language use situation.

A number of studies offered evidence that the iBT elicits performance of the academic language ability (construct) it is designed to assess. The results of a factor analysis, based on a pilot group of iBT test takers, provided empirical evidence that the test items and different parts of the test functioned as intended (Sawaki, Sticker & Oranje 2008). The Sawaki et al. (2008) study also offered support for the four-part design and the use of integrated speaking and writing tasks in the test.

Several studies investigated how examinees’ spoken and written discourse varied on the two types (independent and integrated) of iBT speaking and writing tasks (Brown et al. 2005; Cumming et al. 2006). In both Brown et al.’s study of spoken characteristics and Cumming et al.’s study of written characteristics of examinees’ constructed responses to iBT tasks, results indicated that the abilities (test constructs) the tasks were designed to measure were indeed represented in the discourse examinees produced and their discourse varied, as expected, by their proficiency level and the task. The results of these investigations also provided support for the decision to include both independent and integrated tasks in the speaking and writing components because the discourse elicited by the two tasks differed and produced a more complete representation of the content domain. Cohen & Upton (2006) investigated the
strategies examinees reported using when responding to prototype reading comprehension tasks. Data revealed that examinees responded to reading tasks in ways the test designers intended and the results offered validity evidence for the Reading section of the iBT. Finally, the developers wanted to determine if the various test items and parts of the new TOEFL corresponded to the test constructs. The results of a factor analysis, based on a pilot group of iBT test takers, provided empirical evidence that the test items and different parts of the test functioned as intended (Sawaki, Sticker & Oranje 2008). Sawaki et al.’s study also offered support for the four-part design and use of integrated speaking and writing tasks in the test.

2.1.2.2 Reliability

Reliability is widely referred to as the consistency of test scores or as the absence of error in test scores (Bachman & Palmer 1996: 19; AERA, APA & NCME 1999). Since the degree of reliability in test scores has significant implications for the validity of test score interpretations, reliability is considered by many measurement specialists to be one kind of validity evidence. Commonly reported reliability coefficients include alternate-form, test-retest, internal consistency, and both intra- and inter-rater coefficients. When reliability evidence supports claims the test as a whole or its parts are assessing the construct of interest, it may be reported as validity evidence.

The reliability evidence available to support the validity argument for the iBT comes from several sources. Lee (2006) reported the results of a study that investigated the performance of examinees on sample speaking tasks that were being developed for possible use in the iBT. Since the Speaking section contained three different task types (independent, integrated listening and speaking, and integrated reading and speaking) as well as three different sources of input to the examinee (visual and aural prompt, scripted aural passage, and reading passage), developers wanted to know how these task types affected score reliabilities on the assessment. Findings from the study were useful in several ways. First, the general results represented empirical evidence used to justify certain test design decisions in the Speaking section and complemented the theoretical rationale used to specify the speaking construct. Additionally, the statistical data revealed that the different speaking tasks were highly correlated; all the obtained coefficients were above the .80 level recommended in the Standards for educational and psychological testing (AERA, APA & NCME 1999) and many were over .90. Other reliability evidence from the study revealed the optimal combination of task types, number of tasks, and number of raters needed to achieve acceptable levels of score reliability on the assessment.

An investigation of the sample writing tasks produced reliability evidence similar to that obtained for the sample speaking tasks and the data justified test design decisions for the Writing section (Lee & Kantor 2005). Results indicated that raters did not appear to have a large impact on the reliability of test scores and that the number and type of test tasks had a greater impact on the dependability of scores. The data gave test developers important information on how different parts of the test functioned and provided empirical evidence to support decisions such as how many and what types of tasks to include in the Speaking and Writing sections of the iBT.
ETS used data from every operational test form of the iBT (administered between September 2005 and December 2006) to produce section and total score reliability estimates and the standard error of measurement (Educational Testing Service 2007b). The total score coefficient estimate was .95 and estimates for the Reading and Listening sections were .86 and .87, respectively. Generalizability analyses were used to estimate reliability for the Speaking (.90) and Writing (.78) sections because this procedure is appropriate given the fact the scoring of the constructed responses in these sections involves a degree of judgment. ETS uses multiple raters to evaluate examinees’ performance on the constructed response tasks in the Speaking and Writing sections and raters’ assessments are monitored on a daily basis by means of the OSN system. These procedures contribute to the consistency of both iBT section scores and total scores.

2.1.2.3 Authenticity and interactiveness

Bachman & Palmer (1996: 23) define AUTHENTICITY ‘as the degree of correspondence of the characteristics of a given language test task to the features of a TLU [target language use] task’. As noted above, iBT developers used empirical data collected by a variety of means (including data from surveys, corpora analyses, discourse analyses, and advanced statistical analyses) to explore the design and content of test tasks. Test tasks simulated language use (listening to lectures and out-of-class campus encounters; reading academic texts on a range of subjects; writing summaries and essays of academic content; and communicating orally in academic and social contexts) in the target language use situation. The COE model of the L2 construct and Cumming et al.’s (2005) survey of a sample of ESOL practitioners provided some of the support for the inclusion of integrated tasks.

Based on an examination of the materials contained in the iBT practice book (Educational Testing Service 2006), this author concluded that the content and tasks are generally authentic. The Listening section included two contexts of language use: classroom lectures that simulated the classroom discourse of a broad sample of disciplines, and conversations that simulated two common campus encounters (faculty office visit and student study session). In the samples included with the CD, speakers incorporated many elements of natural discourse into their scripts (e.g., repetitions, false starts, pauses, misspeaks and self-corrections). Reading section passages included selections on topics that represented a broad range of disciplines (zoology, geology, mass media, business, psychology, and political science), and the texts appeared to be comparable to those found in introductory college courses. The integrated speaking and writing tasks represent the same general degree of authenticity found in the content of the Listening and Reading sections.

INTERACTIVENESS is the expected extent of involvement of the examinees’ knowledge and interest and their communicative language strategies in accomplishing a test task (Bachman & Palmer 1996). New item types enhance the interactiveness of each section of the iBT. There are APPLYING PRAGMATIC KNOWLEDGE questions (Listening), READING TO LEARN and PARAPHRASING questions (Reading), and INTEGRATED tasks that require the use of multiple aspects of academic language ability (Speaking and Writing sections). These tasks expand the language knowledge, language functions, and strategies examinees are likely to engage during the test.
2.1.2.4 Impact and practicality

Impact refers to the consequences of using the test and the scores it produces (Bachman & Palmer 1996). These consequences can be positive or negative and they can affect individuals, educational programs, and society. From the outset, ETS considered the impact of the iBT on examinees, test score users, and English teaching practices. One of the principal goals of the iBT project was to improve the positive washback of the new test. Domain analyses and new task types, including the addition of a Speaking component, increase the congruency between what iBT examinees do on the test and the kind of language use that occurs in real academic contexts. The publisher is also interested in the effect of the iBT on language teaching and learning. In order to investigate this aspect of the test, Wall & Horák (2006) began a long-term, multi-stage study prior to the introduction of the iBT that collected data on the pedagogical practices in a sample of ten TOEFL test preparation programs operating in six countries in Europe. These data will serve as a baseline for subsequent comparisons of the pedagogical practices and curriculum content expected to emerge in response to the new test.

Bachman & Palmer (1996) use the term practicality to describe the adequacy of the available resources for the design, development, use, and ongoing evaluation of the test. While these are crucial considerations for anyone who develops a test, publishers of high-stakes, large-scale proficiency assessments must also consider such practical implications as the cost of their test; availability of test dates and sites; time required to take the assessment; time (and cost) required to process and disseminate the results; and perceptions about the test. It is clear from a review of iBT validation activities that ETS carefully considered the practicality of the iBT. One example is the choice of test method (computer delivered) and task design (lack of an interlocutor) used in the Speaking section. The test method and task design made it more practical to collect speech samples and rate them efficiently and consistently.

2.1.3 Commentary

Research on the iBT reflects many of the trends in language testing and assessment. For example, iBT developers used a theoretical conceptualization of the construct of interest (communicative L2 ability in an academic context) and empirical evidence to build a case for the interpretations of and decisions made with iBT scores. Messick (1989) advocated this approach to validation and it has influenced how measurement specialists and test developers conduct their work. While current validation activities focus on building a persuasive validity argument for the interpretation and use of test scores, the process can be organized in different ways. In building the validity argument for the iBT, developers drew on the work of Kane (1992, 2002) and others to determine the structure of the argument and the kinds of evidence to include. The evidence reported by iBT developers includes all the categories of validity evidence (i.e., test content, response processes, internal structure, relations of other variables, and consequences of testing) advocated in the Standards (AERA, APA & NCME 1999).

Researchers utilized multi-method approaches to collect and analyze information on the iBT. In addition to employing survey methods (questionnaires and interviews) and qualitative discourse analysis procedures, developers used advanced statistical procedures
(e.g., generalizability theory, factor analysis, multifaceted Rasch measurement) and corpora analyses to explore the effects of various factors on test performance. By using multiple data collection and analysis procedures, developers obtained more information on more aspects of the iBT than would have been possible with a more restricted range of research tools. The results provided insights into task generalizability, task difficulty, and authenticity considerations. Moreover, recent advances in technology expanded the kinds of technical analysis procedures that could be conducted in the development of the iBT, and these advances also contributed to improvements in the content and administration of the test.

Test developers considered the consequences of test use and sought to maximize the positive consequences associated with the iBT. This is exemplified in a variety of test developer actions from following recognized standards for test validation activities to providing test users information that facilitates test score interpretations and use. ETS has a longstanding practice of codifying the standards and practices it applies to test validation activities (Educational Testing Service 2003a, 2002), and ETS has aligned its practices with those advocated in the Code of fair testing practices in education (Joint Committee on Testing Practices 2004) and the Standards (AERA, APA & NCME 1999).

Clearly, the iBT is an improvement over previous versions of TOEFL. It offers better representation of the L2 construct in general and the English for academic purposes construct in particular. The new test relies less on discrete-point test items and multiple-choice responses, and it utilizes integrated, performance-based tasks that more closely approximate the language use that occurs in academic contexts. The iBT also uses longer samples of discourse in the Reading and Listening sections than the PBT or CBT, and these changes enhance the authenticity and content validity. With the addition of a Speaking component, the iBT provides score users a more comprehensive assessment of examinees’ English proficiency. Test score users are also likely to benefit from iBT validation activities that produced performance descriptions (Educational Testing Service 2004) and work that relates iBT scores to the language ability levels of the CEFR (Tannenbaum & Wylie 2005).

2.2 IELTS

The IELTS (International English Language Testing System) is a collaborative endeavor in which the University of Cambridge ESOL Examinations (C-ESOL), British Council, and International Development Program of Australian Universities and Colleges (IDP: IELTS Australia) share responsibility for researching, administering, and revising the assessment. As with the case of the TOEFL, the development and validation activities related to the IELTS illustrate how high-stakes assessments continually evolve in response to research, theoretical developments, and the expressed needs of stakeholders. The first IELTS appeared in 1989, and it was, in fact, a revised form of the English Language Testing Service (ELTS) that was introduced in 1980. Moreover, the ELTS superseded an earlier measure (the English Proficiency Test Battery) designed to assess the English language ability of international students applying to universities in the UK (Talyor & Falvey 2007).

Data collected during initial IELTS validation activities led developers to modify the test design and administration procedures and to introduce a revised version of the test in 1995.
Following the launch of the 1995 version, the IELTS partners undertook additional validation activities and the data gleaned from these investigations led to subsequent revisions of the Speaking component in 2001 and the Writing component in 2005 (Talyor & Falvey 2007). In fact, between 1995 and 2004 the IELTS partners sponsored over 55 studies related to the measure (Cambridge ESOL 2004). Many of these studies were reported in professional meetings, published in journals, and posted on the Cambridge ESOL website.

It is the paper-based version of IELTS that is discussed here and summarized in Table 2. The information is based on the *IELTS handbook* (Cambridge ESOL 2003), *IELTS specimen materials* (Cambridge ESOL 2005), and the 2005 CD, which contains examples of the listening task input and sample responses to the speaking task. A computer-based version of the test (CB-IELTS) was released in May 2005. The CB-IELTS is not a computer-adaptive test, and the Speaking section is delivered in the same manner as the paper-based IELTS, using a face-to-face format and an interviewer. Examinees who elect to take the CB-IELTS have the option of handwriting their responses to the Writing section tasks or composing them on the computer.

### 2.2.1 Test purpose and method

The *IELTS handbook* (Cambridge ESOL 2003) states the test is ‘designed to assess the language ability of candidates who need to study or work where English is the language of communication’ (p. 2). There are two forms of the test: an academic form designed for use by examinees applying for admission to post-secondary institutions where English is the medium of instruction, and a general form designed for use by those wanting to pursue secondary education, work, training, or permanent residence (immigration) in English speaking countries. C-ESOL reports that in 2003 over 475,000 examinees took the IELTS and 76% of them took the academic form (Blackhurst 2004:18), and information on the IELTS website indicates that 1,000,000 examinees took the test during a recent 12-month period (http://ielts.org). Clearly, the number of IELTS test takers is growing and the impact of the test is significant.

IELTS consists of four sections, Listening, Reading, Writing, and Speaking, and it takes approximately four hours to complete. Table 2 summarizes selected characteristics of the test method. All examinees take the same Listening and Speaking components, and then complete either the Academic or the General Training (GT) Reading and Writing sections. Examinees must complete the Listening, Reading, and Writing sections during a single session. They have the option of taking the Speaking section on the same day or up to seven days prior to or seven days after completing the other three components. Test center clerical staff score examinee’s responses to listening and reading tasks and certified raters score examinees’ writing and speaking responses. All test results are forwarded to C-ESOL for review and analysis. Examinees are notified of their results by the test center within two weeks.

The GT version of the IELTS follows the same format and procedures as the Academic form, including the same number of items (40), scoring procedures, and time allocation (60 minutes). However, the sample reading passages are taken from more general sources.
## Table 2  Selected test method characteristics of IELTS.

<table>
<thead>
<tr>
<th>Component</th>
<th>Input</th>
<th>Response</th>
<th>Time</th>
<th>Evaluation</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Listening</strong></td>
<td>6 2–4 minute monologues and dialogues heard once</td>
<td>Select single correct response (3–5 choices)</td>
<td>30 minutes</td>
<td>Number of correct responses (1 point for each)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Select both correct responses (5–6 choices)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Complete charts, tables, and gaps with 1–3 words, letters, or numbers</td>
<td></td>
<td>Reported as band-level descriptors on scale of 1–9</td>
</tr>
<tr>
<td><strong>Reading</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>(Academic)</strong></td>
<td>3 600–900-word passages</td>
<td>Select single correct response (3–10 choices)</td>
<td>60 minutes</td>
<td>Number of correct responses (1 point for each)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>13–14 questions per passage</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>(GT)</strong></td>
<td>6 200–900-word passages</td>
<td>Select single correct response (3–12 choices)</td>
<td>60 minutes</td>
<td></td>
</tr>
<tr>
<td></td>
<td>From 4–5 questions (short texts) to 8–10 questions (long texts)</td>
<td>Complete table and gaps with 1–3 words, letters, or numbers</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
| Writing | 2 tasks | 60 minutes | Level of performance as judged by trained rater using scale (1–9) and assessment criteria (task fulfillment, response content, coherence and cohesion, vocabulary, and grammar)
|---|---|---|---|
| (Academic) Task 1: written prompt and chart/diagram/graph | 150-word report to professor | (20 minutes) | Task 2 weighted more than Task 1
| Task 2: written prompt | 250-word argument- or issue-oriented essay | (40 minutes) | Penalty for not fulfilling word-length minimum
| (GT) Task 1: written prompt | 150-word letter | (20 minutes) | Reported as band-level descriptors on scale of 1–9
| Task 2: written prompt | 250-word opinion essay | (40 minutes) |
| Speaking | 3-part oral interview | 11–14 minutes | Level of performance as judged by a trained rater using a scale (1–9) and assessment criteria (fluency and coherence, vocabulary, grammar, and pronunciation)
| Interview: interviewer posed questions | Respond to interviewer’s questions | (4–5 minutes) | Reported as band-level descriptors on scale of 1–9
| Extended response: interviewer’s instructions and follow-up questions and written input (task card) | 1 minute to prepare and 1–2 minutes to respond to task card topic and content prompts | (3–4 minutes) |
| Discussion: interviewer’s questions on Part 2 topic | Respond to interviewer’s questions | (4–5 minutes) |
(newspaper advertisements, public notices, promotional materials, and general magazines) whereas the passages in the Academic version are taken from books and periodicals. GT examinees respond to two writing tasks, although the prompts differ slightly from those for the Academic form.

2.2.2 Test justification

C-ESOL operates research and development programs to support its assessments. The activities and findings of these programs are disseminated through volumes in the University of Cambridge Local Examination Syndicate (UCLES)/Cambridge University Press (CUP) Studies in language testing series, the quarterly publication Research Notes (Cambridge ESOL 2007c), and other scholarly publications. Test development and validation activities at C-ESOL are organized into a multi-phase process that includes (1) initial planning and consultation, (2) development, (3) validation, (4) implementation, and (5) operation (Falvey & Shaw 2006). The process is expected to yield assessments that are suitable for their intended purposes and able to satisfy the four qualities (validity, reliability, impact, and practicality) the publisher considers paramount in judging the usefulness of its tests.

Evidence available to support the validity argument for the IELTS comes from multiple sources and investigations. For example, researchers examined (a) the congruence between test tasks and the construct of interest, (b) the target language use situation, (c) the content and format of the test, (d) the characteristics of assessors’ attitudes toward and ratings of constructed responses, (e) task functioning, including whether any tasks were biased toward examinees based on selected examinee characteristics (ethnicity, gender, L1), and (f) the comparability of test versions, including differences between performance on the computer-based and paper-based versions.

2.2.2.1 Construct validity

In early 2008, C-ESOL was engaged in developing an approach to collecting evidence and constructing validity arguments that could be used with all its ESOL exams (Weir & Shaw 2005). The effort was being informed by Weir’s (2005b) socio-cognitive framework, consisting of five elements (context validity, theory-based validity, scoring validity, consequential validity, and criterion-related validity) and covering three dimensions (test taker characteristics, task response, and score) (Weir 2005b: 46–47). At present, the IELTS does not include an explicit theoretical rationale to support the interpretation and use of test scores and one must infer the conceptualizations underlying the test constructs.

Alderson (2000) analyzed the IELTS Reading section and concluded the reading construct was defined in terms of the ability to engage in a set of effective and efficient reading behaviors related to a variety of tasks. As Clapham’s (1996) research on IELTS reading passages revealed, the relationship between reading task performance and text characteristics is complex and affected by test taker characteristics. In terms of the Listening section, one can infer from the tasks and discourse samples that the listening construct includes the ability
to comprehend the main points and important details conveyed in conversations and oral presentations on a variety of topics related to social, academic, and work situations. The writing construct can be inferred by considering how examinees are expected to demonstrate their ability to perform various communicative functions in writing and the assessment criteria used to evaluate the quality of the performance. IELTS writing performance is rated on the basis of task fulfillment, comprehensibility, fluency, grammaticality, and vocabulary. The speaking construct underlying the IELTS assumes competent language users produce comprehensible and fluent language use under time constraints by applying their language knowledge in socially appropriate ways based on the context and purposes of the interaction (Taylor & Falvey 2007).

IELTS developers explored aspects of the task design and content of the two tasks in the Writing component of the Academic IELTS. An item difficulty analysis conducted on Task 1 revealed performances (as measured by the scores awarded and the linguistic characteristics exhibited in responses to the prompts) did not vary significantly due to the prompt (O’Loughlin & Wigglesworth 2007). This suggests that the task functioned as intended and provides empirical support for task design and content decisions. A study by Moore & Morton (2007) investigated how closely Task 2 approximates the writing required in university. The researchers examined a sample of university writing assignments from two universities and interviewed a subset of the faculty who submitted assignments. Results from the task analysis of the target language use domain provided content validity evidence for the Writing component. Kennedy & Thorp (2007) analyzed IELTS examinees’ responses to Task 2. These data informed the development of IELTS band scales and provided evidence for the claim the IELTS elicits the kind of language ability it is intended to assess.

2.2.2.2 Reliability

Chalhoub-Deville & Turner (2000) noted the limited availability of reliability evidence on the IELTS and called for more research into reliability issues related to the test. Since the publication of their article, studies investigating item, alternate forms, and rater reliability have been undertaken and the results published. For example, statistical analyses conducted on the Listening and Reading sections administered in 2004 yielded internal consistency reliability estimates that were from .83 to .91 for the Listening section and averaged .89 across all versions of the test (Cambridge ESOL 2006). The estimates were nearly the same for the Reading section (.83 to .90 and an average of .86 across all versions). An investigation of the comparability of scores on the computer-based and paper-based versions of the test found high correlations between the scores on the two tests based on item difficulty analyses conducted on the Listening and Reading sections (Green & Maycock 2004).

O’Sullivan (2005) reviewed the IELTS and noted that reliability estimates for the Writing and Speaking components were somewhat low. C-ESOL published inter- and intra-rater reliability evidence for the 2004 versions of the Writing and Speaking components and
all the reported obtained reliability correlations exceeded .80 (Cambridge ESOL 2006). The IELTS uses a single rating approach to assessing examinees’ writing and speaking performance, and the publisher contends that adequate reliability evidence exists to justify use of this scoring procedure (Taylor & Falvey 2007). All the reliability correlations reported here for the IELTS meet the minimum thresholds generally advocated by the measurement community, although, in some instances, the reliability data presented in Research Notes may not include enough detail to make judgments about the design or results of the investigations.

2.2.2.3 Authenticity and interactiveness

Careful analysis of the target language use situation helps test developers select the most appropriate content and test tasks to include in an assessment (Douglas 2000; Thighe 2007). The Academic Reading and Writing components of the IELTS have been informed by target language use analyses, but similar analyses have not been conducted for the GT components. Since scores on the GT are being used to make significant decisions that affect an examinee’s ability to practice a profession and immigrate to certain countries, investigations of the target language use domain for the GT version are warranted.

One particular noteworthy feature of the IELTS is the use of an interview format and an interlocutor in the Speaking component. This choice of test method is more authentic than a semi-direct assessment method such as that employed in the iBT. The use of extended passages in the Academic IELTS Reading section also enhances authenticity because longer passages more closely approximate what examinees will encounter in the target language use situation. In the Listening section, examinees hear several British accents as well as American and Australian varieties of English. The four samples on the 2005 CD were scripted and delivered at slightly slower than normal rates, and the samples in parts 1, 2, and 4 represented fairly natural spoken English. This was less true in the case of the conversation among classmates that was depicted in part 3. The speakers’ attempts to include selected features of spoken discourse such as false starts, pauses, fillers, and intonation seemed forced and much less authentic than the other samples. Tasks 1 and 2 of the Academic Writing component employ several common academic writing genres (summary and essay), and there is empirical support for the design and use of these tasks in the IELTS (Mayor et al. 2007; Moore & Morton 2007).

IELTS tasks present examinees with opportunities to engage different kinds of knowledge (linguistic, topical, pragmatic), language skills (listening, reading, writing, and speaking), and metacognitive strategies (monitoring and planning). For example, the interview conducted in the Speaking component requires the examinee to participate in a dynamic conversational exchange with an interlocutor and to share personal information and preferences. Examinees are likely to be motivated and have positive reactions to tasks that are personalized in this way. Additionally, both the Writing and the Speaking components require examinees to use a broad range of knowledge and display multiple aspects of communicative competence in processing the input and formulating their responses to assessment tasks.
2.2.2.4 Impact and practicality

C-ESOL has explored some of the consequences associated with the use of the IELTS. One major initiative was the IELTS impact study. This was a multi-phase project that collected data on different aspects of the test including profiles and performances of IELTS examinees, stakeholders’ attitudes toward the IELTS, and the IELTS' washback on teaching materials and classroom practices (Saville & Hawkey 2004; Hawkey 2006). Findings from the initiative influenced test design decisions, produced evidence of positive consequences of test use, and supported the claim that stakeholders view the IELTS as an acceptable measure of examinees’ English language proficiency.

As a member of the Association of Language Testers in Europe (ALTE), C-ESOL endorses the standards for test quality and fairness advocated in the ALTE Code of practice (ALTE 2001). Two recent initiatives illustrate how C-ESOL test developers are responding to two fairness concerns: sensitivity to test content and appropriate accommodations for examinees with special needs. One project seeks to minimize cultural bias in the content and language used in the publisher’s tests and another project responds to the special needs of examinees by offering them certain accommodations such as additional time, enlarged print, or Braille versions of the IELTS (Gutteridge 2006; Murray 2007).

For more than a decade, the IELTS partners have pursued an active research agenda related to the test and the results have informed changes in test content, design, and scoring procedures. These activities are evidence of the publisher’s support for the IELTS program and commitment to a continuous process of validating the interpretations and use of IELTS scores.

2.2.3 Commentary

IELTS developers organized the validation process into systematic steps that obtained evidence on four important test usefulness qualities (validity, reliability, impact, and practicality) to support the interpretation and use of IELTS scores. As illustrated in the case of the TOEFL, IELTS developers applied multi-method approaches to investigating the test, and they used the findings to improve the measure. When research revealed that rating procedures for the IELTS Writing and Speaking components could be enhanced, developers modified the assessment criteria, scale descriptions, and rater training procedures to improve score dependability on these components of the test. The most compelling validity arguments combine theoretical and empirical evidence to support test score interpretations and use, and the IELTS argument would benefit from more explicit theoretical evidence.

Technological developments have led to several recent innovations in the IELTS program. C-ESOL has begun using the Internet to deliver a web-based Professional Support Network (PSN) that enhances the training and monitoring of raters and expedites the dissemination of test information (Crow & Hubbard 2006). The widespread availability of and familiarity with computers justified the recent introduction of a computer-delivered IELTS at selected test sites. Examinees at these sites now have a choice of taking the computer-delivered or paper IELTS, and if they elect the CB IELTS, they can choose to hand write or word process their responses to writing tasks. The rationale for offering examinees a choice is based on
the assumption that they will choose the option that optimizes their language performance. Preliminary studies conducted by C-ESOL researchers suggest the computer-based version is comparable to the paper-based version (Blackhurst 2005; Maycock & Green 2005; Green & Maycock 2004).

2.3 FCE

Although the FCE (First Certificate in English) first appeared in 1975, it was not an entirely new test but rather a redesigned version of the Lower Certificate in English that was introduced in 1939. A shorter version of the FCE debuted in 1995 following nearly a decade of research and validation activities. In 2003, C-ESOL began a major research and development program on a group of exams that included the FCE. These validation activities led to changes in test specifications and procedures for the version of the FCE scheduled for release in December 2008.

2.3.1 Test purpose and method

The test is intended to assess ‘everyday written and spoken English at an upper intermediate-level . . . for people who want to use English for work or study purposes’ (Cambridge ESOL 2007a). Each component (Reading, Writing, Use of English, Listening, and Speaking) is referred to as a ‘paper’. Examinees take the Reading, Writing, and Use of English papers on the same date, and complete the Listening and Speaking components on a subsequent date within a specified period. The latest version of the FCE takes a little less than four hours to complete (an hour less than the version it supersedes). Test developers reduced the length of the exam by including fewer items in the Reading and Use of English Papers and by having examinees compose shorter responses to one of the writing tasks. Other changes to the test method included eliminating the error correction task type from the Use of English Paper and adding written prompts to the input examinees receive in the second and third tasks of the Speaking component. Table 3 summarizes important test method features of the FCE.

2.3.2 Test justification

C-ESOL completed a number of studies on the FCE in recent years and the results informed and supported revisions to the 2008 FCE (Barker et al. 2007). Validation activities focused on construct definition; internal consistency reliability estimates for the Reading and Listening components; rater-reliability evidence for the Writing and Speaking components; and authenticity, impact, and fairness considerations. The process included obtaining the opinions of key stakeholder constituencies (examinees, assessors, educators, and school administrators), reviewing the constructs and specifications of the test, and investigating the effect of changes to the test method (e.g., using fewer tasks, modifying the nature of the input and length
Table 3  Selected test method characteristics of FCE.

<table>
<thead>
<tr>
<th>Component</th>
<th>Input</th>
<th>Response</th>
<th>Time</th>
<th>Evaluation*</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reading</td>
<td>3 passages of 550–700 words 30 questions</td>
<td>Select correct response from set of 4–8 choices</td>
<td>60 minutes</td>
<td>Number of correct responses (1–2 points each)</td>
</tr>
<tr>
<td>Writing</td>
<td>2 tasks</td>
<td>Compose 120–150-word letter or email</td>
<td>80 minutes</td>
<td>Level of performance as judged by a trained rater using scale (0–5) and assessment criteria (content, organization and cohesion, register and format, language use, ability to inform)</td>
</tr>
<tr>
<td></td>
<td>Task 1: written prompt</td>
<td>Select topic from 5 choices and compose 120–150-word response in 1 of 6 genres</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Task 2: written prompt</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Use of English</td>
<td>3 gapped passages</td>
<td>Select correct response from set of 4–12 choices</td>
<td>45 minutes</td>
<td>Number of correct responses (1–2 points each)</td>
</tr>
<tr>
<td></td>
<td>Set of 8 questions (each containing an initial and a gapped sentence)</td>
<td>Fill in the gap with correct form and/or up to 5 words</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Listening</td>
<td>Multiple 30-second – 3-minute monologues and dialogues</td>
<td>Select correct response from 3–7 choices</td>
<td>40 minutes</td>
<td>Number of correct responses (1 point each)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Fill in the gap with the correct information (1–3 words)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
### Table 3 (Continued)

<table>
<thead>
<tr>
<th>Component</th>
<th>Input</th>
<th>Response</th>
<th>Time</th>
<th>Evaluation*</th>
</tr>
</thead>
<tbody>
<tr>
<td>Speaking**</td>
<td>Part 1: interlocutor’s questions</td>
<td>Respond to interlocutor’s questions</td>
<td>14 minutes</td>
<td>Level of performance as judged by trained raters (2) using different scales (holistic/analytic) and assessment criteria (overall performance/grammar and vocabulary, discourse management, pronunciation, and communication)</td>
</tr>
<tr>
<td></td>
<td>Part 2: interlocutor’s instructions and prompt, two photographs, written question above photographs, and other examinee’s discourse</td>
<td>Produce 1-minute response to input and 20-second response to other examinee’s discourse</td>
<td>(4 minutes)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Part 3: interlocutor’s instructions, several visuals, and other examinee’s discourse</td>
<td>Interact with other examinee, discuss visuals, and express opinions</td>
<td>(3 minutes)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Part 4: interlocutor’s questions on topic addressed in Part 3</td>
<td>Respond to questions and engage in discussion</td>
<td>(4 minutes)</td>
<td></td>
</tr>
</tbody>
</table>

*All five FCE components are weighted equally. Performance on the FCE is reported on a scale of 0–100 and letter grades (A–E) are assigned to score ranges (e.g., scores from 80–100 are designated Grade A and scores of 54 or below are designated Grade E).

**The FCE speaking component uses a paired format in which two examinees and two assessors are present.
of expected responses, and reducing the time allotted). Data collected from these activities constitute some of the evidence available to support a validity argument for the FCE.

2.3.2.1 Construct validity

Previous FCE specimen materials did not explicitly define the construct(s) of interest. The FCE handbook for teachers (Cambridge ESOL 2007b: 3) states that L2 proficiency consists of the ‘language user’s overall communicative ability’ and it includes descriptions of the specific language skills (reading, writing, lexico-grammar, listening, and speaking constructs) assessed by the test. Weir & Shaw (2005) reported that a socio-cognitive framework (Weir 2005b) was used to formulate the writing construct.

FCE developers used a variety of research procedures to investigate the changes proposed for the Writing, Listening, and Speaking components, and they included key stakeholders in the process. Researchers obtained stakeholders’ responses to the proposed task designs by using questionnaires, verbal protocols, observations, and focus group discussions. The results influenced and justified various task design decisions (Barker et al. 2007; Fried-Booth 2007). Corpora analyses were conducted on the content of the Reading component (Hughes 2006) and they were used to design the speaking tasks (Barker 2006). These recent FCE validation activities reveal how test developers are using qualitative and quantitative procedures to investigate tests, and how they are integrating stakeholders into the test development process. Test developers must be careful not to rely on content validity evidence to establish the construct validity of a test. If content validity evidence is not connected to theory-based evidence, it can increase the risk of underrepresentation of the test construct, unrepresentative samples of the content domain, or inclusion of construct irrelevant knowledge, skills, or abilities in the test (Embretson 2007).

2.3.2.2 Reliability

A study commissioned by the publisher and conducted by Bachman et al. (1995) in the late 1980s found generally low internal consistency reliability coefficients for the FCE Listening, Reading, and Use of English Papers and no evidence of rater reliability estimates for the Writing and Speaking Papers (Spolsky 1995: 341). Buck (2001) believed that increasing the number of items and adopting a four-option response format would improve reliability in the Listening component.

FCE developers focused on the issues raised in the Bachman et al. study and explored ways to minimize the unmotivated variation in examinees’ scores that was attributable to test content, tasks, different test forms, and raters’ scoring of writing and speaking performance. They addressed some of these issues by (a) restricting choice, as in the case of the second FCE writing task where examinees select one of five options, (b) providing more specific guidelines and additional training for raters, (c) collecting and monitoring rater reliability data, and (d) collecting and analyzing task difficulty by means of advanced statistical procedures that explored the interaction between task performance and different aspects of the test method (Weir 2005b).
These efforts appear to have contributed to the results of recently published reliability estimates. Purpura (2004) reported internal consistency estimates of .88–.93 for the Use of English Papers administered between 1999 and 2003. Cooze & Shaw (2007) conducted a study of raters’ scorings of examinees’ performance on prototype tasks for Task 1 in the Writing Paper and obtained inter-rater reliability estimates of around .70. Galaczi (2005) reported inter-rater reliability coefficients of .79–.85 for a sample of raters’ scorings of Speaking Papers administered in 2003. Rater reliability data on the Speaking component is particularly important because Orr (2002) found that raters’ scorings of FCE speaking performances varied and the raters did not follow or attend to the evaluation criteria. He recommended improving the rating scales and rater training. It is also important to demonstrate that examinees’ scores are not adversely affected by the test method (a paired format in which two examinees and two assessors are present) used to elicit speaking performance. Research suggests that interlocutor performance varies and the interlocutor and the other examinee influence performance on this component (Lazaraton 1996; O’Sullivan 2004, 2002).

### 2.3.2.3 Authenticity and interactiveness

FCE writing tasks replicate many of the general purpose writing tasks examinees can expect to encounter in academic and employment contexts, and input to examinees establishes an audience and purpose for their responses. The speaking tasks also simulate real-world language use in that examinees must extemporaneously converse with an interviewer, speak on a topic, negotiate with an interlocutor, and engage in a discussion with multiple participants.

According to the *FCE handbook for teachers*, the Listening section samples represent ‘a variety of voices, styles of delivery, and accents’ (Cambridge ESOL 2007b: 53). Although Buck (2001: 230) concluded that input to examinees was ‘reasonably authentic,’ he thought it should include ‘more hesitations, more spoken grammar and more spoken vocabulary choice’. Canagarajah (2006) suggested the Speaking component of the FCE could be made more sensitive to examinees’ use of different varieties of spoken English if one of the raters and one of the examinees came from different speech communities. While this arrangement might lead to more authentic interaction, it would be important to investigate the effect of this test method facet on examinees’ scores before adopting it.

### 2.3.2.4 Impact and practicality

FCE developers considered the impact of the test on examinees and sought to maximize the positive consequences associated with the measure. For example, the *FCE handbook for teachers* (Cambridge ESOL 2007b) contains extensive instructions on how teachers can prepare students to perform optimally on test tasks. Moreover, FCE tasks correspond to the kinds of language learning and use examinees experience in their classes and this contributes to positive washback. Results of a recent survey of a sample of FCE stakeholders found attitudes toward the test to be largely positive (Barker et al. 2007).

In principle, test developers want to enhance the positive consequences of an assessment and achieve an adequate balance among the different test usefulness qualities, but in practice
this can be quite challenging. This is illustrated in the test method facets in Part 1 of the FCE Listening component. In Part 1, examinees hear the listening scripts, questions, and three possible answers twice and read the questions and three choices. It can be argued that listeners typically do not hear the content twice in real-world language use situations and therefore this test method facet reduces authenticity. On the other hand, if examinees hear the input twice, it may reduce anxiety, evoke a positive reaction, more fully engage their listening competence, and better demonstrate the test construct. Thus, this feature of the test method enhances the interactiveness quality. When FCE researchers surveyed a sample of examinees about their attitudes toward this test method characteristic, three quarters of the respondents indicated they liked it, although modifying it would reduce the length of time required to complete the test and examinees would likely view this as a positive consequence, too (Fried-Booth 2007). In consideration of the expressed preferences of test users, FCE developers retained this test method feature and found other ways to reduce the length of the test.

As was noted in the case of the IELTS, C-ESOL conducts an ongoing program of research and development related to the FCE. The results of these validation activities inform test design decisions and demonstrate a desire to promote positive test use consequences. It is clear from the information available on the FCE revision project that C-ESOL possesses the resources to deliver, monitor, and revise the test as needed.

2.3.3 Commentary

Recent validation activities appear to have addressed some of the previous concerns with construct definition and the adequacy of reliability evidence for the FCE. In particular, reliability evidence now exists for three components of the test (Writing, Use of English, and Speaking) and the reported coefficients meet the standards generally advocated for high-stakes assessments.

Efforts to improve the quality of the reliability evidence do not appear to have come at the cost of continuing to place a high priority on the use of authentic test tasks. This is exemplified in the Writing component where examinees are given the topic in the case of Task 1 and have a choice of topics in Task 2. Giving examinees a choice has potential implications for the reliability of scores; however, FCE developers have limited the number of choices, standardized administration procedures, and striven to make tasks comparable in difficulty in order to mitigate the effect of unmotivated variation in performance on the tasks (Weir 2005b). Some of the additional reliability evidence available to support use of the revised FCE includes revisions to test specifications, improvements in rater training, and the data obtained from piloting test tasks.

2.4 TOEIC

The TOEIC (Test of English for International Communication) was originally commissioned by the Japanese government and designed for use in making recruitment, placement, and
training decisions with Japanese workers. Since its introduction in 1979, the number of examinees taking TOEIC has expanded steadily and in 2007 4.9 million registrants in 92 countries took the exam (Mary Enright, personal communication, 28 May 2008). There were two parts to the 1979 test (Listening and Reading) and each contained 100 multiple-choice questions. The test content was based on what test developers expected examinees would encounter in workplace settings where English was used for general commercial purposes. A revised TOEIC Listening and Reading test was introduced in 2005 and it continues to be a paper-based assessment. In late 2006, TOEIC Speaking and Writing tests were launched. These are computer-delivered assessments that are administered separately and at different times than the TOEIC Listening and Reading test.

2.4.1 Test purpose and method

TOEIC is designed to assess ‘the everyday English skills of people working in an international environment . . . and scores indicate how well people can communicate in English with others in business, commerce, and industry’ (Educational Testing Service 2007c: 2). Tables 4.1 and 4.2 summarize key test method characteristics of the TOEIC. ETS uses the Internet and OSN system to collect and distribute TOEIC examinees’ speaking and writing samples to raters and to train and monitor rater scoring.

2.4.2 Test justification

To date, validation activities conducted on the TOEIC have been more limited than those reported for the TOEFL, but this will probably change in the coming years because the audience for the test is growing rapidly and expanding beyond Asia. Given the paucity of published research on the measure, the increasing heterogeneity of examinees, and the introduction of TOEIC Speaking and Writing tests, additional validity and reliability evidence is required to support the test score interpretations and use of the test with a global audience.

According to the publisher, the 2007 TOEIC ‘reflects global business communication styles and emphasizes authentic language contexts’ and requires examinees ‘to use multiple strategies and abilities to comprehend and connect information’ (Educational Testing Service 2007d). If so, the revised Listening and Reading sections will have enhanced authenticity and examinees will find these components now offer greater interactivity. These changes also address some of the perceived limitations in previous versions of the TOEIC. Douglas (2000) reviewed an earlier version of the Reading component and concluded that it was a ‘good example of a well-constructed norm-referenced traditional multiple-choice test task, with no doubt high reliability, but extremely limited in the inferences it will allow about language knowledge’ (p. 236).

In a review of the Listening component, Buck (2001: 210) stated that it was a ‘classic’ example of the multiple-choice type language tests associated with the U.S. psychometric tradition, and he expressed some concerns with the test. These included the lack of variety
Table 4.1  Selected test method characteristics of the TOEIC Listening and Reading test.

<table>
<thead>
<tr>
<th>Component</th>
<th>Input</th>
<th>Response</th>
<th>Time</th>
<th>Evaluation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Listening</td>
<td>Statements, questions, conversations, and talks heard once</td>
<td>Select correct response from 3–4 choices</td>
<td>45 minutes</td>
<td>Number of correct responses (1 point each)</td>
</tr>
<tr>
<td>Part 1</td>
<td>Visual (photograph) followed by spoken instructions and 4 descriptions</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Part 2</td>
<td>30 spoken questions/statements each followed by 3 possible responses</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Parts 3 &amp; 4</td>
<td>20 30–45 second conversations or monologues followed by 3 questions each (spoken input)</td>
<td>Each question and 4 possible responses (written input)</td>
<td>Reported on scale of 5–495</td>
<td></td>
</tr>
<tr>
<td>Reading</td>
<td>Multiple texts (25–150 words)</td>
<td>Select correct response from 4 choices</td>
<td>75 minutes</td>
<td>Number of correct responses (1 point each)</td>
</tr>
<tr>
<td></td>
<td>Some texts in part 7 (reading comprehension) are inter-related</td>
<td></td>
<td></td>
<td>Reported on scale of 5–495</td>
</tr>
<tr>
<td>Test (Task)</td>
<td>Input</td>
<td>Response</td>
<td>Time</td>
<td>Evaluation</td>
</tr>
<tr>
<td>------------------</td>
<td>---------------------------------</td>
<td>-----------------------------------------------</td>
<td>--------</td>
<td>-----------------------------------------------------------------------------</td>
</tr>
<tr>
<td>Speaking</td>
<td>Written instructions and prompts; photographs</td>
<td>0–45 seconds to prepare and 15–60 seconds to respond</td>
<td>20 minutes</td>
<td>Level of performance as judged by trained raters and reported on scale of 0–200</td>
</tr>
<tr>
<td>(Read aloud)</td>
<td>Written instructions and text</td>
<td>45 seconds to prepare and 45 seconds to respond</td>
<td></td>
<td>Raters use scale (0–3) and assessment criteria (pronunciation and intonation and stress)</td>
</tr>
<tr>
<td>(Describe picture)</td>
<td>Written instructions and photograph</td>
<td>30 seconds to prepare and 45 seconds to respond</td>
<td></td>
<td>Raters use scale (0–3) and assessment criteria (pronunciation, intonation and stress, grammar, vocabulary, and cohesion)</td>
</tr>
<tr>
<td>(Respond to question)</td>
<td>Written instructions and written and spoken prompts</td>
<td>No time to prepare and 15–30 seconds to respond</td>
<td></td>
<td>Raters use scale (0–3) and assessment criteria (pronunciation, intonation and stress, grammar, vocabulary, cohesion, relevance and completeness of content)</td>
</tr>
<tr>
<td>(Respond to question and visual)</td>
<td>Written instructions, written and spoken prompts, and visual</td>
<td>30 seconds to read written input before spoken prompt begins and 15–30 seconds to respond</td>
<td></td>
<td>Raters use scale (0–3) and assessment criteria (pronunciation, intonation and stress, grammar, vocabulary, cohesion, relevance and completeness of content)</td>
</tr>
<tr>
<td>(Solve problem)</td>
<td>Written instructions and written and spoken prompt</td>
<td>30 seconds to prepare and 60 seconds to respond</td>
<td></td>
<td>Raters use scale (0–5) and assessment criteria (pronunciation, intonation and stress, grammar, vocabulary, cohesion, relevance and completeness of content)</td>
</tr>
<tr>
<td>Task Type</td>
<td>Instructions</td>
<td>Time Limit</td>
<td>Assessment Details</td>
<td></td>
</tr>
<tr>
<td>------------------------</td>
<td>----------------------------------------</td>
<td>------------</td>
<td>------------------------------------------------------------------------------------</td>
<td></td>
</tr>
<tr>
<td>(Express opinion)</td>
<td>Written instructions and spoken prompt</td>
<td>15 seconds</td>
<td>60 seconds to prepare and 60 seconds to respond Raters use scale (0–5) and assessment criteria (pronunciation, intonation and stress, grammar, vocabulary, cohesion, relevance and completeness of content)</td>
<td></td>
</tr>
<tr>
<td>Writing</td>
<td>Compose 5 sentences, 2 emails, and opinion essay</td>
<td>60 minutes</td>
<td>Level of performance as judged by trained raters and reported on a scale of 0–200</td>
<td></td>
</tr>
<tr>
<td>(Construct sentence)</td>
<td>Construct single sentence based on photograph and two specified words or phrases in prompt</td>
<td>(8 minutes)</td>
<td>Raters use a scale (0–3) and assessment criteria (grammar and relevance of sentence to prompt)</td>
<td></td>
</tr>
<tr>
<td>(Respond to request for information)</td>
<td>Compose email based on prompt</td>
<td>(20 minutes)</td>
<td>Raters use scale (0–4) and assessment criteria (sentence quality and variety, vocabulary, and organization)</td>
<td></td>
</tr>
<tr>
<td>(Express opinion essay)</td>
<td>Compose essay based on prompt</td>
<td>(30 minutes)</td>
<td>Raters use scale (0–5) and assessment criteria (support for opinion, grammar, vocabulary, and organization)</td>
<td></td>
</tr>
</tbody>
</table>
in test tasks, limitations in the nature of the input to examinees (that is, there was ‘very little fast speech’ and an absence of some features typically associated with spoken discourse), and the elicitation of responses that emphasized narrow aspects of the listening construct (p. 216).

2.4.2.1 Construct validity

TOEIC validation activities began when the first operational version of the test was launched in Japan (Woodford 1982). Investigators found relatively high degrees of correspondence between Japanese examinees’ performance on the TOEIC Listening and Reading components and other direct assessments of listening and reading ability (Woodford 1982; Wilson 1999). Initial validation activities included computing correlations between the TOEIC Reading section and two different components of the TOEFL (the Reading Comprehension and Vocabulary Section and the Structure and Written Expression Section). Wilson (2000) reported that the obtained correlations were .85 and .87, respectively, and that subsequent studies conducted with examinees from other ethnolinguistic backgrounds have yielded similar findings. Hence, there are several forms of criterion-related validity evidence reported for the original Listening and Reading components of the TOEIC. Additional criterion-related validity evidence was collected by correlating a sample of examinees’ scores on the TOEIC and self-assessments of their ability to perform selected tasks in English (e.g., read office memoranda, comprehend spoken instructions, order a product by telephone, and make a written list of items). The reported correlations were modest and ranged from .40 to .50 (Educational Testing Service 2007a).

It is not clear from the research and practice materials available on the ETS website if the TOEIC was designed on the basis of a theoretical framework, but recent work has related the TOEIC Listening and Reading test to the CEFR (Tannenbaum & Wylie 2005). The current construct of interest appears to be general communicative language ability, and one can infer it is comprised of listening, reading, speaking, and writing skills and the examinee’s ability to use these skills for social and occupational purposes in everyday life. The handbook (Educational Testing Service 2007a) states that the Listening and Reading sections were redesigned based on the results of ETS research, but no details are offered. There is limited validity evidence available on the latest TOEIC, but the iBT project demonstrated the kinds of validation activities ETS conducts on its assessments. If the publisher undertook similar investigations of the TOEIC, the results would strengthen the validity argument for the test.

2.4.2.2 Reliability

Zhang (2006) examined the performance of a large sample of Japanese and Korean examinees that took the same form of the TOEIC in 2003. The obtained reliability estimates for the 2003 sample were very similar to those found in the 1979 validation study. They were .91–.93 for Listening, .92–.93 for Reading, and .95 and .96 for the Listening and Reading total scores, respectively. Moreover, the publisher reported that internal reliability estimates for the
Listening and Reading section scores were .90 or above for all forms of the test (Educational Testing Service 2007a).

As of early 2008, TOEIC publications did not contain reliability evidence for the new TOEIC Speaking and Writing tests. ETS contends its use of a ten-step process enhances score reliability for these components. The process includes carefully selecting and training assessors, utilizing scoring rubrics and detailed instructions, supervising and monitoring rater performance, and analyzing raters’ performances prior to disseminating examinees’ scores (Educational Testing Service 2007e). While these activities are commendable and promote score dependability, additional reliability evidence should be collected and used to support the validity argument for the new TOEIC.

### 2.4.2.3 Authenticity

The 2007 TOEIC included updated input (pictures and texts) in the Listening and Reading test and this contributes to authenticity. Listening input is made more international by including three varieties of English (Australian, British, and North American). Texts in the Reading Section samples (email messages and a letter from a business establishment) correspond to the target language use situation. In the TOEIC Speaking and Writing tests, sample tasks required examinees to fulfill communicative functions (e.g., problem solution and expressing an opinion) that are appropriate for the target audience and consistent with the test purpose. Nonetheless, the authenticity of the TOEIC Speaking test is constrained by the test method. Speaking to a computer does not engage important aspects of an examinee’s communicative competence, including the co-construction of discourse that is thought to be fundamental to interactional competence (Fulcher 2003).

### 2.4.2.4 Impact and practicality

Clearly, the impact of the TOEIC is growing as more examinees take the test and as use of it expands beyond Asia. There is evidence that TOEIC developers considered how to maximize the positive impact of the test. The addition of Speaking and Writing tests extends the skills covered by TOEIC and offers examinees and score users a more comprehensive assessment of English language ability. Those examinees who prepare to take the full TOEIC battery will likely experience more positive washback than those who prepare to take a single TOEIC test.

As noted above, ETS follows test development procedures designed to ensure its assessments meet established quality and fairness standards. In the case of the TOEIC, staff researchers and external experts analyzed the content of the TOEIC and strove to eliminate any bias in test items that would make it unsuitable for use in a global context (Educational Testing Service 2003b). Item writers avoided vocabulary, idiomatic expressions, or other features of language use that were limited to a particular variety of English and emphasized the use of situations, professional activities, and content that would be recognized and relevant to examinees from a broad range of countries.
2.4.3 Commentary

Zhang (2006) wanted to determine the effect of Japanese and Korean examinees’ language backgrounds on scores, but the results of the study were inconclusive. Because the TOEIC was designed initially for and used extensively with Japanese examinees, it is important to continue to investigate the measure and to determine that it is not systematically biased against examinees from other language backgrounds. As noted above, item reliability evidence was reported for the 2003 Listening and Reading test and the standard error of measurement is reported in the 2007 examinee manual. The reporting of these data is consistent with prevailing professional practice and it is advocated in the Standards (AERA, APA & NCME 1999). That said, there is limited reliability and validity evidence available for the new TOEIC Speaking and Writing tests and additional evidence is needed to support test score interpretations and use. The most significant change to the TOEIC is the availability of Speaking and Writing tests. In view of the significance of speaking and writing ability in current conceptualizations of the L2 construct, the availability of this option makes the TOEIC a more comprehensive measure of L2 ability than in the past. Many examinees and score users will perceive this as a positive development.

3. Conclusion

Authoritative publications like the ALTE (2001) Code of practice and the Standards (AERA, APA & NCME 1999) are intended to promote sound test development and proper test use practices. Both test developers and users have responsibility for validating the proposed score interpretations and uses of high-stakes assessments. Current approaches to the validation process entail collecting information from multiple sources and using both theoretical and empirical evidence to construct a validity argument. Experience suggests that the strongest validity arguments emerge from a validation process that is systematic, comprehensive, logical, and ongoing. In judging the case for a particular test, one should bear in mind that developers are balancing various test usefulness considerations and working with the available evidence. Since the process is an ongoing effort, test users can expect developers to refine their validity arguments as more evidence becomes available and as developments in language testing occur. The evaluation of high-stakes tests is less about arriving at a summative decision and is more about considering the adequacy of the overall evidence and the relative strength of the argument for the test in its context of use. If stakeholders are to participate more fully in the validation process and judge validity arguments, test developers will need to make the notion of validity and the validation activities they undertake more transparent and comprehensible to a general audience. As demonstrated by the discussion of the tests considered in this article, there appears to be some progress on this front.

What is the way forward? Many assessment specialists acknowledge the paucity of practical frameworks to guide test developers in prioritizing and conducting their work (Kane 2004; Moss 2007). Some of the validation frameworks proposed by applied linguists have begun to address this gap, but more research and professional consensus will be required
How should the theoretical and empirical (quantitative and qualitative) evidence be prioritized and integrated into a coherent validity argument? How much evidence and what kinds of evidence are needed to support the validity arguments for a large-scale, high-stakes test? The pursuit of answers to these questions will shape the direction of future developments in validation.

Research suggests characteristics of the test method, characteristics of the examinees, and the processes and strategies examinees use in response to test tasks affect performance on language tests (Bachman 2000; Leung & Lewkowicz 2006). Advanced statistical procedures (factor-analysis, structural equation modeling, and generalizability theory) are facilitating investigations of construct validity, score dependability, and sources of bias, but the interactions and relationships among the many variables that affect test performance are complex and not fully understood. Performance-based tasks offer the prospect of improving the authenticity and interactiveness qualities of high-stakes tests, but researchers must further investigate and resolve concerns (authenticity, task difficulty, and generalizability) with this type of task for the field to move forward.

At the 2001 meeting of the Association of Language Testers in Europe, Alderson (2004) made some prescient predictions on what to expect in the years ahead. He speculated there would be more research into washback in particular and the impact of tests in general, including the ethical and political implications associated with test development and use. He submitted that the Common European Framework of Reference would continue to evolve and affect language teaching and learning and that developments in computer technology would improve and expand the availability of classroom-based assessments. He believed models of the L2 construct would continue to be investigated and new research methods would expand our understanding of the complexities associated with test methods and test performance.

As we near the end of the first decade of the new millennium, it is evident that developments in language testing are proceeding as Alderson predicted. In the decade ahead, we can expect more progress on the issues identified by Alderson. The research methods used to investigate language assessments have expanded dramatically in recent years. Test developers are now applying qualitative procedures such as discourse analysis, survey methods, and ethnographies as well as increasingly sophisticated statistical procedures to investigate tests (Lumley & Brown 2005). Future developments related to validation frameworks, progress in understanding performance-based tasks, and applications of an expanded array of sophisticated research tools and methods will contribute to improvements in the quality and fairness of high-stakes assessments over the next decade.

Acknowledgements

I wish to express my appreciation to Graeme Porte for his support throughout the process and to thank the editorial board members, four anonymous reviewers, and Mary Enright for the valuable feedback they provided during the development of this review. Any remaining limitations in the manuscript are wholly my own.
List of acronyms

AERA  American Educational Research Association
ALTE  Association of Language Testers in Europe
APA   American Psychological Association
CB-IELTS  Computer-based International English Language Testing System
CBT   Computer-based Test (Test of English as a Foreign Language)
CEFR  Common European Framework of Reference
C-ESOL University of Cambridge English to Speakers of Other Languages Examinations
ELTS  English Language Testing Service
ESOL  English to speakers of other languages
ETS   Educational Testing Service
FCE   First Certificate in English
iBT   Internet-based Test (Test of English as a Foreign Language)
IELTS International English Language Testing System
NCME National Council on Measurement in Education
OSN   Online Scoring Network
PBT   Paper-based Test (Test of English as a Foreign Language)
PSN   Professional Support Network
TESOL Teachers of English to Speakers of Other Languages
TOEFL Test of English as a Foreign Language
TOEIC Test of English for International Communication
UCLES University of Cambridge Local Examination Syndicate

References


**Stephen Stoynoff** is Professor of English and Director of Graduate Studies in English at Minnesota State University, Mankato, in the United States, where he teaches courses in second language assessment, second language literacy, and research methods. He is co-editor (with Carol Chapelle) of *ESOL tests and testing* (TESOL, 2005) and was editor of *TESOL Journal* in 1998–2003.