Lang. Teach. (2012), **45.2**, 234–249 © Cambridge University Press 2012
doi:10.1017/S026144481100053X

Thinking Allowed

Research agenda: Priorities for future research in second language assessment

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

1. Introduction

In a recent state-of-the-art (SoA) article (Stoynoff 2009), I reviewed some of the trends in language assessment research and considered them in light of validation activities associated with four widely used international measures of L2 English ability. This Thinking Allowed article presents an opportunity to revisit the four broad areas of L2 assessment research (conceptualizations of the L2 construct, validation theory and practice, the application of technology to language assessment, and the consequences of assessment) discussed in the previous SoA and to propose tasks I believe will promote further advances in L2 assessment. Of course, the research tasks I suggest represent a personal stance and readers are encouraged to consider additional perspectives, including those expressed by Bachman (2000), Chalhoub-Deville & Deville (2005), McNamara & Roever (2006), Shaw & Weir (2007), and Stansfield (2008). Moreover, readers will find useful descriptions of current research approaches to investigating L2 assessments in Lumley & Brown (2005), Weir (2005a), Chapelle, Enright & Jamieson (2008), Lazaraton (2008), and Xi (2008).

Research in language assessment has expanded greatly in recent years and there are many resources available to those searching for research topics and methods. Noteworthy resources include the journals *Language Assessment Quarterly* and *Language Testing*; research reports posted on the websites of Cambridge ESOL, Educational Testing Service and other major test developers; and titles in *Studies in Language Testing* (Cambridge University Press) and the *Cambridge Language Assessment Series*. Since much of the cutting edge research in language assessment today is conducted or sponsored by major test development centers, the studies undertaken in conjunction with these entities can be exemplars, offering insights not only on what to investigate in terms of an assessment but also how to investigate it. While much of the published research has focused on measures of L2 English ability, the issues and research designs also apply to studies of L2 assessments developed for a wide range of local contexts and assessment purposes.

https://doi.org/10.1017/S026144481100053X Published online by Cambridge University Press

2. Four broad strands of L2 assessment research

2.1 Representations of the L2 construct

Theoretical representations of communicative L2 ability provide a reasoned basis and useful starting point from which to define the construct of interest, design appropriate assessment tasks, and justify the score-based inferences and uses of tests. Bachman (2007) characterizes prevailing theoretical representations of the L2 construct as 'ability-focused, task-focused, and interaction-focused', and he submits that while they 'are not mutually exclusive', it may not be possible to reconcile them – at least theoretically – in part because they are based on different 'values and assumptions' (p. 41). Nonetheless, construct-related research that investigates the nature of L2 ability constitutes some of the most fundamental research to be pursued in language assessment.

For several decades, the consensus view has been that communicative L2 ability is multicomponential, consisting of a set of highly interrelated subcompetencies that interact in a particular language use situation (Purpura 2008), although it is unclear exactly how many factors are involved, how they are interrelated, and how they affect language performance. Therefore, empirical studies that substantiate and elucidate the subcompetences believed to underlie the nature of L2 ability not only contribute valuable evidence to support theoretical representations of the construct but also provide crucial support for the design and use of particular L2 assessments.

The investigations of both the paper-based (PBT) and internet-based (iBT) versions of the Test of English as a Foreign Language™ (TOEFL®) are a case in point. Based on a review of studies that explored the factor structure of the TOEFL, Sawaki, Stricker & Oranje (2009) concluded there was empirical support for the multicomponential perspective, although the number and combination of factors identified across the different studies varied. Moreover, in a confirmatory factor analysis of a prototype of the iBT, Sawaki et al. found the factor structure of the new test to be best represented by a higher-order factor model that included a general factor (L2 ability) and four group factors, each related to one of the parts of the test (Listening, Reading, Speaking, and Writing). In another study of the same sample of test-takers, Stricker & Rock (2008) found the same factor structure reported by Sawaki et al. across subgroups, despite differences in the study participants' number of years of classroom instruction in English and their first language. Because these studies were conducted on a prototype of the iBT TOEFL, studies of an operational version of the test using a representative sample of the iBT test-taker population are warranted. Additionally, if similar studies were conducted on other L2 assessments, the results would expand the available evidence for or against the multicomponential model and possibly further explicate the contributions of various factors to the construct.

Of course, factor-analytic studies represent only one form of quantitative evidence and one approach to investigating language assessments. Test developers recognize the need to utilize multiple forms of evidence, collected throughout the test development and implementation process. While quantitative procedures have often been emphasized in test validation, qualitative procedures are increasingly being used as well.

https://doi.org/10.1017/S026144481100053X Published online by Cambridge University Press

Survey methods (questionnaires and interviews), discourse and conversation analysis, verbal report (or recall) protocol analysis, and observations are among some of the most frequently cited qualitative procedures, and researchers have employed them in a number of recent studies of international assessments of L2 English ability. For example, Lazaraton (2002) used a conversation analysis procedure to collect and analyze interlocutors' discourse on the Key English Test (KET) and the Certificate in Advanced English (CAE), and Galaczi (2008) used the procedure to investigate the oral discourse produced by paired test-takers on the speaking component of the First Certificate in English (FCE). TOEFL researchers used discourse analysis procedures and verbal report analysis protocols to explore the academic language ability construct and the knowledge, processes, and strategies used by iBT test takers (Brown et al. 2005; Cohen & Upton 2006; Cumming et al. 2006). Orr (2002) employed the verbal protocol procedure to investigate rater behaviors on the speaking performances of FCE test takers. In a recent International English Language Testing System (IELTS) study, Wigglesworth & Elder (2010) used discourse analysis and survey data (questionnaires and group interviews) to gain insights into the effect of planning time on test takers' speaking performance in Part 2 of the IELTS speaking section. Although each of these studies investigated assessments of L2 English ability, the research designs are relevant to investigations of other L2 assessments.

In an effort to better understand the nature of L2 ability, researchers have examined a number of the underlying constructs thought to contribute to it. Among these are the knowledge, skills, and cognitive processes L2 users engage when listening, reading, speaking, and writing in a language use situation or when responding to language assessments. For example, Swain et al. (2009) investigated strategic competence, which is considered to be a key aspect of communicative language ability, by examining test takers' reported strategy use in response to tasks on a modified version of the speaking section of the iBT. The researchers used a questionnaire and an immediate recall procedure (following the completion of each of six assessment tasks) to collect data from test takers, and the results of the study offer some empirical support for the role of cognitive processes in the L2 speaking construct. Weir et al. (2007) also used an immediate recall procedure following each task on a practice version of the IELTS, to investigate the cognitive processes and strategies test takers reported using in response to tasks in the reading section.

The results of these studies offer empirical evidence of the processes and strategies elicited by certain assessment tasks, and provide additional insights into the L2 reading and speaking ability constructs. They also represent approaches one can undertake with other L2 assessments. In fact, without the results of investigations of more L2 assessments conducted in more contexts, it will be difficult to attain a complete understanding of the contributions of the factors involved in interactive L2 communication, or how they interact.

Research task 1:
Conduct studies that replicate or adapt exemplary research designs used to investigate international measures of L2 English ability (e.g. those reported in sources cited in the Introduction and studies highlighted in section 2.1) and use the results to increase understanding of the L2 construct and support the score-based inferences and uses of other L2 assessments.

https://doi.org/10.1017/S026144481100053X Published online by Cambridge University Press

In the past decade, some scholars have conceptualized L2 ability as a set of hierarchically arranged communicative functions, situations, topics, and acts and the relative competence displayed by L2 users across a range of proficiency levels. This perspective is exemplified in the Common European Framework of Reference (CEFR). The CEFR emerged from a European initiative intended to facilitate TRANSPARENCY and COHERENCE among those engaged in language education by creating a common scheme for communicating, organizing, reflecting on, monitoring, and promoting L2 learning, teaching, and assessment (North 2008: 22). However, its influence now extends far beyond Europe. Given the global impact of the CEFR and some of the concerns raised by assessment specialists, it is important to continue to investigate the framework and to accumulate construct-related support for its use (for critiques of the CEFR, see, for example, Fulcher 2004, Weir 2005b, Hulstijn 2011, and Little 2011).

One concern with the CEFR has to do with the proficiency level descriptions and the adequacy of the illustrative scale descriptors for some competences in the framework. To address these limitations, North (2008) advocates undertaking investigations in local contexts using a process that collects and categorizes samples of performances and test items across proficiency levels, compares the data to CEFR descriptor scales, and uses the results to support descriptors or justify modifications to them based on empirical evidence.

Since ability in one component of the L2 construct can differ from the knowledge and skill displayed in another and because competence in different components (such as linguistic versus strategic competence) can vary in terms of both the rate of L2 development and the level of attainment, it is important to understand more fully the contribution of various components to overall L2 ability and how they interact in different language use contexts. Research that addresses the following task will contribute additional insights into the nature of L2 ability and may provide construct-related evidence for the CEFR.

Research task 2:
Investigate the competences, knowledge, and processes language users engage at various levels of L2 proficiency, from beginning levels through advanced levels.
Investigate the language performance of test takers on locally developed L2 assessments across proficiency levels and compare the results to CEFR scale descriptors.

CEFR designers intended the framework to be language- and context-neutral and capable of being adapted to meet local circumstances. While a number of studies have reported on its use within a European context, there are limited published accounts of how it has been applied or adapted for use in other regions. Wu & Wu (2010) describe a project undertaken in Taiwan that sought to align a locally developed L2 assessment with the CEFR, based on the approach advocated in a draft document distributed by the Council of Europe (2003). Yet without more case studies of how the CEFR has been used in other contexts and how locally developed L2 assessments have been effectively aligned with the framework, it will be difficult to empirically support claims that the CEFR is truly capable of being adapted to contexts anywhere in the world. A related but somewhat different issue has to do with how much modification can reasonably occur in applications of the CEFR to local circumstances before L2 standards cease to be comparable or assessments cease to be aligned with the

https://doi.org/10.1017/S026144481100053X Published online by Cambridge University Press

CEFR. Studies that respond to the following research task will help address these issues and also provide some of the validity evidence needed to justify the use of local L2 assessments.

Research task 3:
Investigate the extent to which the CEFR can be applied or adapted to local circumstances and yet remain aligned with and comparable to the standards and assessments of other CEFR users.

The context of language use figures prominently in current representations of the L2 construct, including the CEFR. Brown (2008) has speculated that the language ability construct may actually be composed of 'a cluster of CONSTRUCTS that vary from setting to setting depending on variations in purposes, approaches, syllabuses, testing methods, constraints, and so on' (p. 288). If this is correct, it is important to conduct studies proposed in research task 4 using frameworks that promote systematic and comprehensive analyses of the test use contexts and facilitate identification of the most salient features in the context. For a detailed description of a context analysis conducted on a local assessment of L2 English ability (EIKEN) used in Japan, see Brown (2008).

Research task 4:
Conduct test use context analyses of significant L2 assessments and determine the most salient features operating in each context.

Given the increased use of performance-based tasks in L2 assessment, it is important to have empirical evidence both to inform the design and to justify use of the tasks. At present we have an incomplete understanding of how task and test taker characteristics affect task performance or how the various factors related to these characteristics interact and contribute to performance (Bachman 2002). Recent research has also revealed the influence of assessors on test takers' performances and some of the factors that appear to affect the awarding of scores on performance-based tasks. Research task 5 proposes investigations that will increase understanding of key issues related to performance-based task assessment. To date, most studies have focused on a few high-stakes tests of L2 English ability, and it is important to explore the use of these tasks in a broader range of L2 assessments and contexts, including their use in large-scale national and international testing programs and in classroom-based assessment.

Research task 5:
Investigate test takers' performances on assessment tasks including the processes and strategies they use to fulfill the tasks.
Investigate the effects of key test taker characteristics (e.g. gender, L1, years of exposure to L2) on task performance.
Investigate the extent to which the task design, including planning time, number of tasks, rating procedures, and number of raters, affects score reliability.
Investigate the extent to which selected factors, such as characteristics of the assessor, test taker, task, rating scheme, and assessor training, affect assessors' rating of task performance.

https://doi.org/10.1017/S026144481100053X Published online by Cambridge University Press

Unfortunately, there are too few accounts of how task characteristics and test taker characteristics affect performance on assessments designed for local assessment purposes. A recent example of research that explores this is Gan's (2010) school-based case study of test takers' performance on a group oral assessment. Gan used a conversation analysis procedure to investigate the discourse that emerged from higher- and lower-scoring groups of L2 learners on a classroom-based group oral assessment. Data from the study revealed features of the interactional discourse produced by groups with different levels of L2 proficiency and provided empirical support for use of a group oral assessment task. See the 2009 special issue of *Language Testing* (26.3) for current approaches to investigating paired and group assessments of L2 performance and Taylor & Wigglesworth's overview of key issues.

In another example of a local study with implications for L2 classroom-based assessment, Matsuno (2009) used multifaceted Rasch measurement to compare self- and peer-ratings of L2 writing performance with teacher ratings. In addition to offering some empirical support for the use of peer assessment of classroom-based writing tasks, the study represents a research design that could be used to investigate assessor effects on other types of performance-based tasks and L2 assessments. Studies that investigate the extent to which L2 learners can function effectively as assessors of L2 performance have implications for efforts to empower learners and to integrate assessment alternatives into L2 classroom practices. Research task 6 seeks to collect empirical evidence to support the use of practical classroom-based task designs and assessment procedures.

Research task 6:
Conduct studies to investigate how the use of self, paired, or group assessment affects L2 performance (including the nature of the oral and written discourse produced by language users).
Conduct studies to investigate how the assessor affects the ratings awarded for L2 performance.

2.2 Validation theory and practice

As Chapelle (1999) observed, there are numerous obstacles to advancing work in validity, including the lack of adequate models for defining the L2 construct, the difficulty of building a validity argument to justify use of an assessment, and the paucity of practical frameworks to help test developers and users translate validation theory into concrete actions that improve assessment practices. Despite these challenges, current standards of professional practice compel test developers to justify (validate) the interpretations and uses of test scores. Much of the recent work related to validation theory in L2 assessment has been informed by Kane (see Kane 2006 and 2010 for descriptions of an argument-based approach to validation), Mislevy and colleagues (Mislevy, Steinberg & Almond 2002), Bachman (2005), and Weir (2005a), among others.

The case study of the TOEFL iBT project compiled by Chapelle et al. (2008) exemplifies the current argument-based approach to validation. In addition to rendering a comprehensive, longitudinal account of the validation activities undertaken to support the design and the

https://doi.org/10.1017/S026144481100053X Published online by Cambridge University Press

interpretive and validity arguments for the new TOEFL test, the case study presents one of the most detailed accounts of L2 test development available. Bachman & Palmer (2010) proffer a framework designed to guide practitioners in the development of L2 assessments and the arguments used to justify the inferences and uses of scores. The ASSESSMENT USE ARGUMENT that emerges from the Bachman & Palmer framework incorporates recent theoretical developments in validation and the pattern of logical reasoning and argument structure of Toulmin (2003). Finally, Weir (2005a) has integrated recent developments in validation theory into an evidence-based approach to test development and validation based on six key considerations (components): TEST TAKER, COGNITIVE VALIDITY, CONTEXT VALIDITY, SCORING VALIDITY, CONSEQUENTIAL VALIDITY, and CRITERION-RELATED VALIDITY. His socio-cognitive framework focuses on the collection of multiple forms of complementary evidence for each of the six components of the framework throughout the test development and implementation process. The accumulated body of validity evidence is used to support claims about the quality and usefulness of the assessment. Chapter 11 of Weir (2005a) presents a number of illustrative applications of the socio-cognitive framework, as well as a variety of quantitative and qualitative procedures available for the investigation of L2 assessments.

Since the frameworks described above are relatively new, there are limited reports of their impact on language assessment practices. Therefore, case studies that examine how these frameworks have been implemented in different contexts and how they have contributed to the validation practices associated with local L2 assessments will establish the extent to which validation theory is capable of promoting improvements in validation practices. Several recently published accounts suggest that these frameworks are affecting the development of local L2 assessments. For example, Pardo-Ballester (2010) reports on how Bachman's (2005) assessment use argument framework was used to construct the validation argument for a local placement test of L2 Spanish listening ability, and Llosa (2011) describes a study that built a validity argument for the interpretation and use of student performances on a classroom-based assessment that was aligned with state-mandated L2 proficiency standards. Llosa's multi-method research design employed confirmatory factor analysis and a verbal protocol procedure to obtain data and determine if a standardized measure of students' L2 proficiency assessed the same construct as a classroom-based assessment that was based on teacher judgments.

The global movement in standards-based L2 teaching, learning, and assessment has led to a need for locally developed standards-based classroom assessments; however, there are limited published case studies of empirically validated standards-based assessments. Thus, studies of how teachers have applied appropriate validation frameworks to the development of standards-based classroom assessments will not only reveal the extent to which existing validation frameworks can be adapted or applied to the development of standards-based classroom assessments but will also establish an empirically based justification for the use of the assessments. As a number of scholars have emphasized, however, the validation of classroom-based assessments may require new ways of looking at key test qualities such as reliability and validity (McKay 2005; Rea-Dickens 2008; Davison & Leung 2009). Research task 7 is intended to determine the impact of validation frameworks on L2 assessment practice and to encourage the validation of locally developed L2 assessments.

https://doi.org/10.1017/S026144481100053X Published online by Cambridge University Press

Research task 7:
Conduct case studies that reveal the extent to which current validation frameworks can be effectively applied to the validation of locally developed L2 assessments.
Undertake studies that report on the development and validation activities associated with L2 standards-based classroom assessments.

Current approaches to validation emphasize the need to collect multiple forms of evidence to justify test use to stakeholders. In order to gauge the impact of validation theory on the professional activities of assessment specialists, Cizek, Rosenberg & Koons (2008) reviewed the test descriptions contained in the 2005 edition of the *Mental measurements yearbook* (Spies & Plake 2005). The researchers concluded that the reported validity information did not conform to current views of validity theory and that consequential validity evidence was essentially absent. A comparable study of the test descriptions of assessments of L2 English ability or other major L2 assessments, as proposed in research task 8, would establish a baseline from which to monitor subsequent progress in the adoption of professional standards relative to validity evidence, and would document the effect of developments in educational measurement on the field of language testing.

Research task 8:
Conduct a study of the validity evidence included in the reports disseminated by major L2 test publishers and determine how closely the information included in the reports conforms to current professional standards and perspectives on validation.

2.3 Technology

The application of computer technology to large-scale assessment offers great promise, but raises a number of concerns as well. One fundamental issue is how the application of computer-assisted language testing (CALT) affects test taker performance and, more narrowly, whether test taker performances on paper-based and computer-based versions of the same language test are comparable. While the narrower question has been investigated in several studies conducted on large-scale assessments of L2 English ability, several assessment specialists have called for more studies that explore the extent to which there might be group differences in performance on computer-mediated assessments (Chapelle & Douglas 2006; Leeson 2006; Douglas & Hegelheimer 2007). Douglas & Hegelheimer (2007) posited that computer familiarity may be a factor for some groups of test takers, and it is incumbent on test developers to collect evidence that test taker characteristics such as gender, first language, geographical region, or socioeconomic status do not lead to significant differences in the degree of familiarity with computer technology, and that this in turn is not a source of unmotivated variation in test performance.

A study conducted by Choi, Kim & Boo (2003) represents a multi-method approach to investigating the comparability of scores on paper- and computer-based versions of the same L2 assessment. The researchers collected empirical evidence of comparability by using a variety of means, including content analysis, correlational analyses, analysis of variance, and confirmatory factor analysis. In addition to providing empirical support for the claim that

https://doi.org/10.1017/S026144481100053X Published online by Cambridge University Press

performances on two versions of the same assessment were comparable, the study represents a research design that can be replicated in other contexts where both computer-based and paper-based versions of the same L2 assessment are used.

Weir et al. (2007) used a multi-method design to investigate the performances of test takers on a paper and a computer-mediated version of IELTS writing tasks. The results produced empirical support for the claim that the test method in this case does not significantly affect assessors' ratings of performance and also supported the claim that test takers engage in similar cognitive processing when responding to paper-based and computer-mediated versions of the same task. Additionally, based on study participants' questionnaire responses, test takers were found to be generally familiar and comfortable with computers. However, given the characteristics and number of participants in the study, the researchers recommended replicating it with a larger and more representative sample of test takers. In another study that explored computer familiarity and the effect of test method on performance, Yu (2010) reported minimal differences in the performances of test takers on an integrated reading and writing task and the test takers' degree of computer familiarity did not appear to affect performance. While the three studies cited above demonstrate some empirical support for the use of computer-mediated versions of tests, they need to be complemented by studies that examine large representative samples of the test taker population for a given assessment. In order to ensure the validity of score inferences for an assessment, it is also important to investigate the effects of all the task types used in an assessment and to determine the extent to which paper and computer-mediated versions elicit the same knowledge, skills, cognitive processing, and oral and written discourse from test takers. Studies undertaken in response to research task 9 will contribute empirical evidence to justify the use of computer-mediated assessments of L2 ability.

Research task 9:
Conduct studies of computer-mediated L2 assessments that utilize representative samples of target test taker populations to investigate the influence of key test taker characteristics (e.g. gender, first language, geographical region, socioeconomic status) on test performance.
Conduct studies that investigate the effects of computer-mediated L2 assessments on test taker performance including the language knowledge, skills, cognitive processing, and discourse elicited by various assessment tasks.

Advances in technology make it possible to include a broader range of tasks in computer-mediated assessments than was possible even a few years ago. Today a variety of media are available to enhance the authenticity of a task, but the media input (e.g. video, audio, and graphics) may also increase the processing demands placed on test takers and possibly compromise their performance. A study by Wagner (2007) investigated how much time test takers spent viewing video input on listening assessment tasks, and he found the participants spent an average of two-thirds of the allotted time watching the video input. In another part of the same study, Wagner (2010) compared test takers' performances on a listening assessment based on whether the task provided video or only audio input, and found that the group that received the video input scored higher and the difference between the two groups was statistically significant. These data offer some preliminary empirical evidence

https://doi.org/10.1017/S026144481100053X Published online by Cambridge University Press

of how test method characteristics and technology-enhanced media input affect test takers' performance, but the issue remains largely uninvestigated. Studies undertaken in response to research task 10 will offer insights into this aspect of computer-mediated assessment.

Research task 10:
Conduct studies that investigate the influence of particular computer-mediated test method characteristics, such as the need to scroll to read the text, the ability to review test items or not, and various media input, on test taker performance.

In a study of computer-mediated assessment, Wise & Kong (2005) sought to investigate test takers' motivation. If test takers are not motivated to put forward their best effort on a test, it undermines test score validity. Empirical investigations of test motivation have typically relied on self-report data collected from test takers during or after the test. In contrast, Wise & Kong used the principle of response time effort (RTE) as an indicator of test takers' motivation and analyzed the amount of time test takers spent on test items. The researchers then combined this quantitative data with self-report data they collected from test takers. This study demonstrates how a multi-method research design using complementary data collection procedures can collect empirical evidence of how test taker motivation and test method characteristics (computer-mediated test) affect performance on an assessment. Because this issue has not been explored with regard to L2 assessments, research task 11 proposes investigating it.

Research task 11:
Conduct an approximate replication of Wise & Kong's study to determine the effect of test takers' motivation and test method characteristics (computer-mediated assessment) on L2 performance.

2.4 Consequences of assessment

The terms WASHBACK and IMPACT are often used to describe studies that investigate the consequences associated with tests and test use. While precise definitions of the terms remain somewhat elusive, the emerging view is that IMPACT is an overarching concept that refers to the broad influences of tests on individuals, educational programs and systems, and society at large. WASHBACK, on the other hand, is more often used to describe the effects of assessment on language teaching and learning (Hamp-Lyons 1997; Hawkey 2006).

Several assessment specialists have proposed conceptual frameworks for investigating the impact of language assessments. Chalhoub-Deville (2009) argues that impact studies need to begin before the language assessment is introduced rather than after the fact, which is commonly the case. She advocates conducting a SOCIAL IMPACT ANALYSIS (SIA) and contends that this will more effectively integrate key stakeholders into the process of considering the consequences of test use and testing practices. The TESTING-CONTEXT ANALYSIS framework proposed by Brown (2008) also emphasizes the need to integrate stakeholders into a process that includes data collection and analyses of test use consequences. Given the limited number

https://doi.org/10.1017/S026144481100053X Published online by Cambridge University Press

of reports on how these frameworks have been implemented, research task 12 proposes that case studies be undertaken to demonstrate the efficacy of these frameworks.

Research task 12:
Conduct a social impact analysis or a testing-context analysis as part of the test validation activities undertaken with an L2 assessment, and report on how the analysis contributed to the consideration of test use consequences.

Few longitudinal impact studies have been conducted on major L2 assessments. Two notable examples are the multi-phase impact studies conducted on the IELTS and iBT tests. The IELTS impact study employed a multi-method approach to collect survey (questionnaire and interview) and observation data from informants in order to determine the influence of the test on stakeholders in the teaching-learning context as well as those beyond it. Data from the three-phase study and from other impact studies commissioned by the IELTS research program have contributed consequential validity evidence to justify use of the test.

iBT researchers also adopted a multi-method approach to investigate the impact of the new TOEFL. They utilized survey methods (questionnaires and interviews) and classroom observations to obtain data from key stakeholders (teachers, students, and program administrators), and used the data to document the classroom teaching and learning practices in a sample of Central and Eastern European contexts prior to the introduction of the test. The data also revealed how test users acquired information about the iBT, how well they understood the new TOEFL, and how they responded to the new features of the assessment (Wall & Horák 2006; Wall & Horák 2008). The third phase of the impact study was designed to collect information on classroom teaching and learning following the introduction of the iBT and to support claims that the test fulfilled its intended purpose. Both the iBT and IELTS impact studies represent designs that can be adopted for longitudinal studies of other L2 assessments.

One aspect of test impact that has not been sufficiently explored is the extent to which test users are able to understand the information disseminated about the test by the test publisher. Based on a review of the test reports distributed by large-scale assessment programs, Goodman & Hambleton (2004) called for more empirical studies of how information about major tests is communicated, to score users and their ability to understand the information. The researchers suggested that think aloud protocols could be used to gain insights into how certain subgroups interpret and understand reports. Test publishers can use the results to improve and tailor the information disseminated to test users based on the characteristics of different test user groups. Studies undertaken in response to research task 13 will provide a form of consequential validity evidence not currently available for L2 assessments.

Research task 13:
Investigate how test information is communicated to stakeholders and determine if test takers and score users are able to understand the information and interpret test results appropriately.

Most of the research on test impact has focused on the effects of major national and international L2 assessments. This has led some to call for more studies that explore the

https://doi.org/10.1017/S026144481100053X Published online by Cambridge University Press

impact of assessment on individual students and classroom learning (Purpura 2009; Cheng & Curtis 2012). In fact, very little is known about how different assessments and assessment practices affect individual test takers' cognitive processing and their acquisition of key L2 competences. Consequently, there is a need for studies that investigate the effects of assessment on students' learning and L2 development; such research will not only provide consequential validity evidence to support the use of specific local L2 assessments but will also foster stronger connections between assessment practices and L2 learning.

Research task 14:
Conduct studies that investigate the effects of assessment on students' learning and L2 development including how it affects learners' cognitive processing and acquisition of various L2 competences.

The language-testing community has devoted considerable attention to both professional standards and ethics. While progress has been made in codifying and promulgating guidelines intended to promote the development of robust tests and the appropriate use of test scores, there is a paucity of empirical research regarding the effect of professional standards on professional practice. Hence, studies undertaken in response to research task 15 will provide preliminary evidence of the degree to which test developers and test users report they are familiar with professional standards and acknowledge the influence of those standards on how they conduct their work (Alderson 2006).

Research task 15:
Survey a sample of test developers and test users and determine the degree to which respondents are aware of relevant professional standards and codes of practice (e.g. those endorsed by the Association of Language Testers in Europe, International Language Testing Association, or the Joint Committee of the American Educational Research Association, American Psychological Association, and National Council on Measurement in Education) and acknowledge their influence on the respondents' professional practice.

TEST FAIRNESS concerns receive major treatment in the standards and codes of practice disseminated by authoritative professional groups in the international language testing community. The centrality of fairness to language assessment has led several researchers to propose frameworks to facilitate its consideration (Kunnan 2004, 2008; Xi 2010). Using the iBT as an example, Xi (2010) demonstrates how a fairness framework can be used to identify, order, and integrate various kinds of empirical evidence into a fairness argument that supports 'the comparability of the score-based interpretations and uses for different relevant groups' (p. 158). Recent articles by Cheng & Qi (2006) and Shih (2008) demonstrate how Kunnan's fairness framework can facilitate consideration of large-scale national assessments of L2 English ability. However, there is a need for reports on how fairness frameworks can be integrated effectively into the entire validation process from the design of the assessment through its implementation and continued use, and research task 16 addresses this.

Research task 16:
Adopt or adapt a test fairness framework and report on how it was used to address fairness concerns and validation activities related to an L2 assessment.

https://doi.org/10.1017/S026144481100053X Published online by Cambridge University Press

Recently, fairness concerns have been raised about large-scale international assessments of L2 English ability. In particular, some assessment specialists believe the tests may be biased in favor of native speaker norms and have called for more empirical investigations to ascertain if this is indeed the case (Chalhoub-Deville & Wigglesworth 2005; Davidson 2006; McNamara & Roever 2006; Hamp-Lyons & Davies 2008). In a study conducted by Hamp-Lyons & Davies (2008) on the Michigan English Language Assessment Battery (MELAB), the researchers sought to determine if there is evidence the MELAB favors English native speaker norms and that the test is biased against certain non-native speakers' English. Because of limitations in the design of the study (sample size, rater characteristics, rater selection, and text selection), the researchers were unable to draw any definitive conclusions and they recommended additional investigations. Moreover, they argued it is important to continue to examine 'whose norms should be imposed in these tests and what the consequences for test takers are if the norm imposed by the test is not the ''normal'' variety accepted in their own society' (p. 26). Taylor (2009) discusses some of the theoretical and practical considerations teachers and testers face with regard to language standards and language variation and the results of a survey of ALTE members on this issue (Taylor 2008). Given the significant consequences of major international assessments of L2 ability and the limited number of empirical studies of this issue, investigations related to research task 17 are warranted.

Research task 17:
Conduct studies to determine if major international L2 tests advantage some groups of test takers over others in terms of the language use norms that are privileged. Investigate the degree of significance for any group differences that are found to exist and determine the impact of the differences on test takers.

3. Conclusion

In this Thinking Allowed article, I have suggested resources for those in search of topics and current approaches to language assessment research. I have also highlighted some recent studies that investigated specific language assessments and addressed issues in each of the four strands covered in this article. To date, most of the research in language assessment has focused on a limited number of large-scale, high-stakes assessments of L2 English ability. While this work has contributed to general developments as well as improvements in specific tests, it is crucial over the next decade to build upon and extend this research to a much broader range of L2 assessments and contexts. Studies conducted in response to the tasks I proposed in this article offer the prospect of not only advancing research in each of the four strands but also enhancing the practices associated with the development and use of various L2 assessments.

Acknowledgement

I wish to express my appreciation to Graeme Porte and four anonymous reviewers for their helpful comments on an initial draft of this article.

https://doi.org/10.1017/S026144481100053X Published online by Cambridge University Press

References

Alderson, J. C. (2006). Bridging the gap between theory and practice? Paper presented at the meeting of the European Association for Language Testing and Assessment. Krakow, Poland.

Bachman, L. F. (2000). Modern language testing at the turn of the century: Assuring that what we count counts. *Language Testing* 17.1, 1–42.

Bachman, L. F. (2002). Some reflections on task-based language performance assessments. *Language Testing* 19.4, 453–476.

Bachman, L. F. (2005). Building and supporting a case for test use. *Language Assessment Quarterly* 2.1, 1–34.

Bachman, L. F. (2007). What is the construct? In J. Fox, M. Wesche, D. Bayliss, L. Cheng, C. Turner & C. Doe (eds.), *Language testing reconsidered*. Ottawa: University of Ottawa Press, 41–71.

Bachman, L. & A. Palmer (2010). *Language assessment in practice*. New York: Oxford University Press.

Brown, A., N. Iwashita, T. McNamara & S. O'Hagan (2005). *An examination of rater orientations and test taker performance on English-for-academic-purposes speaking tasks* (TOEFL Monograph 29). Princeton, NJ: Educational Testing Service.

Brown, J. D. (2008). Testing-context analysis: Assessment is just another part of language curriculum development. *Language Assessment Quarterly* 5.4, 275–312.

Chalhoub-Deville, M. (2009). The intersection of test impact, validation, and educational reform policy. *Annual Review of Applied Linguistics* 29, 118–131.

Chalhoub-Deville, M. & C. Deville (2005). A look back at and forward to what language testers measure. In E. Hinkel (ed.), *Handbook of research in second language teaching and learning*. Mahwah, NJ: Erlbaum, 815–831.

Chalhoub-Deville, M. & G. Wigglesworth (2005). Rater judgment and English language speaking proficiency. *World Englishes* 24.3, 383–391.

Chapelle, C. A. (1999). Validity in language assessment. *Annual Review of Applied Linguistics* 19, 254–272.

Chapelle, C. A. & D. Douglas (2006). *Assessing language through computer technology*. Cambridge: Cambridge University Press.

Chapelle, C. A., M. K. Enright & J. M. Jamieson (eds.) (2008). *Building a validity argument for the Test of English as a Foreign Language*. New York: Routledge.

Cheng, L. & A. Curtis (2012). Test impact and washback: Implications for teaching and learning. In C. Coombe, P. Davidson, B. O'Sullivan & S. Stoynoff (eds.), *The Cambridge guide to second language assessment*. Cambridge: Cambridge University Press, 89–95.

Cheng, L. & L. Qi (2006). Description and examination of the National Matriculation English Test. *Language Assessment Quarterly* 3.1, 53–70.

Choi, I., K. Kim & J. Boo (2003). Comparability of a paper-based language test and a computer-based language test. *Language Testing* 20.3, 295–320.

Cizek, G., S. Rosenberg & H. Koons (2008). Sources of validity evidence for educational and psychological tests. *Educational and Psychological Measurement* 68.3, 397–412.

Cohen, A. & T. Upton (2006). *Strategies in responding to the new TOEFL reading tasks* (TOEFL Monograph 33). Princeton, NJ: Educational Testing Service.

Council of Europe (2003). *Relating language examinations to the CEFR. Manual preliminary pilot version*. Strasbourg: Council of Europe.

Cumming, A., R. Kantor, K. Baba, U. Erdosy & M. James (2006). *Analysis of discourse features and verification of scoring levels for independent and integrated prototype written tasks for the next generation TOEFL* (TOEFL Monograph 30). Princeton, NJ: Educational Testing Service.

Davidson, F. (2006). World Englishes and test construction. In B. Kachru, Y. Kachru & C. Nelson (eds.), *The handbook of world Englishes*. Oxford: Blackwell, 709–730.

Davison, C. & C. Leung (2009). Current issues in English language teacher-based assessment. *TESOL Quarterly* 43.3, 393–415.

Douglas, D. & V. Hegelheimer (2007). Assessing language using computer technology. *Annual Review of Applied Linguistics* 27, 115–132.

Fulcher, G. (2004). Deluded by artifices? The Common European Framework and harmonization. *Language Assessment Quarterly* 1.4, 253–266.

Galaczi, E. (2008). Peer-peer interaction in a speaking test: The case of the First Certificate in English examination. *Language Assessment Quarterly* 5.2, 89–119.

https://doi.org/10.1017/S026144481100053X Published online by Cambridge University Press

Gan, Z. (2010). Interaction in group oral assessment: A case study of higher- and lower-scoring students. *Language Testing* 27.4, 585–602.
Goodman, D. & R. Hambleton (2004). Student test score reports and interpretive guides: Review of current practices and suggestions for future research. *Applied Measurement in Education* 17.2, 145–220.
Hamp-Lyons, L. (1997). Washback, impact, and validity: Ethical concerns. *Language Testing* 14.3, 295–303.
Hamp-Lyons, L. & A. Davies (2008). The English of English tests: Bias revisited. *World Englishes* 27.1, 26–39.
Hawkey, R. (2006). *Impact theory and practice*. Cambridge: UCLES/Cambridge University Press.
Hulstijn, J. (2011). Language proficiency in native and non-native speakers: An agenda for research and suggestions for second-language assessment. *Language Assessment Quarterly* 8.3, 229–249.
Kane, M. (2006). *Validation*. In R. Brennan (ed.), *Educational measurement* (4th edn). Westport, CT: Praeger, 16–64.
Kane, M. (2010). Validity and fairness. *Language Testing* 27.2, 177–182.
Kunnan, A. (2004). Test fairness. In M. Milanovic & C. Weir (eds.), *European language testing in a global context*. Cambridge: UCLES/Cambridge University Press, 27–48.
Kunnan, A. (2008). Towards a model of test evaluation: Using the test fairness and test context frameworks. In L. Taylor & C. Weir (eds.), *Multilingualism and assessment*. Cambridge: UCLES/Cambridge University Press, 229–251.
Lazaraton, A. (2002). A qualitative approach to the validation of oral language tests. Cambridge: UCLES/Cambridge University Press.
Lazaraton, A. (2008). Utilizing qualitative methods for assessment. In E. Shohamy & N. Hornberger (eds.), *Encyclopedia of language and education* (2nd edn), *Language testing and assessment* (volume 7). New York: Springer, 197–209.
Leeson, H. (2006). The mode effect: A literature review of human and technological issues in computerized testing. *International Journal of Testing* 6.1, 1–24.
Llosa, L. (2011). Standards-based classroom assessment of English proficiency: A review of issues, current developments, and future directions for research. *Language Testing* 28.3, 367–382.
Little, D. (2011). The Common European Framework of Reference for Languages: A research agenda. *Language Teaching* 44.3, 381–393.
Lumley, T. & A. Brown (2005). Research methods in language testing. In E. Hinkel (ed.) *Handbook of research in second language teaching and learning*. Mahwah, NJ: Erlbaum, 833–855.
Matsuno, S. (2009). Self-, pair-, and teacher-assessments in Japanese university EFL writing classrooms. *Language Testing* 26.1, 75–100.
McKay, P. (2005). Research into the assessment of school-age language learners. *Annual Review of Applied Linguistics* 25, 243–263.
McNamara, T. & C. Roever (2006). *Language testing: The social dimension*. Ann Arbor, MI: Blackwell.
Mislevy, R., L. Steinberg & R. Almond (2002). Design and analysis in task-based language assessment. *Language Testing* 19.4, 477–496.
North, B. (2008). CEFR levels and descriptor scales. In L. Taylor & C. Weir (eds.), *Multilingualism and assessment*. Cambridge: UCLES/Cambridge University Press, 21–66.
Orr, M. (2002). The FCE speaking test: Using rater reports to help interpret scores. *System* 30.2, 143–154.
Pardo-Ballester, C. (2010). The validity argument of a Web-based Spanish listening exam: Test usefulness evaluation. *Language Assessment Quarterly* 7.2, 137–159.
Purpura, J. (2008). Assessing communicative language ability: Models and their components. In E. Shohamy & N. Hornberger (eds.), *Encyclopedia of language and education* (2nd edn), *Language testing and assessment* (volume 7). New York: Springer, 53–68.
Purpura, J. (2009).The impact of large-scale and classroom-based language assessments on the individual. In L. Taylor & C. Weir (eds.), *Language testing matters*. Cambridge: UCLES/Cambridge University Press, 301–325.
Rea-Dickens, P. (2008). Classroom-based assessment. In E. Shohamy & N. Hornberger (eds.), *Encyclopedia of language and education* (2nd edn), *Language testing and assessment* (volume 7). New York: Springer, 257–271.
Sawaki, Y., L. Stricker & A. Oranje (2009). Factor structure of the TOEFL Internet-based test. *Language Testing* 26.1, 5–30.

https://doi.org/10.1017/S026144481100053X Published online by Cambridge University Press

Shaw, S. & C. Weir (2007). *Examining writing: Research and practice in assessing second language writing*. Cambridge: UCLES/Cambridge University Press.
Shih, C. (2008). The General English Proficiency Test. *Language Assessment Quarterly* 5.1, 63–76.
Spies, R. & B. Plake (eds.) (2005). *The sixteenth mental measurements yearbook*. Lincoln, NE: Buros Institute of Mental Measurements.
Stansfield, C. (2008). Where we have been and where we should go. *Language Testing* 25.3, 311–326.
Stoynoff, S. (2009). Recent developments in language assessment and the case of four large-scale tests of ESOL ability. *Language Teaching* 42.1, 1–40.
Stricker, L. & D. Rock (2008). *Factor structure of the TOEFL Internet-based test across subgroups* (TOEFL iBT Research Report 07). Princeton, NJ: Educational Testing Service.
Swain, M., L. Huang, K. Barkaoui, L. Brooks & S. Lapkin (2009). *The speaking section of the TOEFL iBT: Test takers' reported strategic behaviors* (TOEFL iBT Research Report 10). Princeton, NJ: Educational Testing Service.
Taylor, L. (2008). Language varieties and their implications for testing and assessment. In L. Taylor & C. Weir (eds.), *Multilingualism and assessment*. Cambridge: UCLES/Cambridge University Press, 276–295.
Taylor, L. (2009). Setting language standards for teaching and assessment: A matter of principle, politics, or prejudice? In L. Taylor & C. Weir (eds.), *Language testing matters*. Cambridge: UCLES/Cambridge University Press, 139–157.
Taylor, L. & G. Wigglesworth (2009). Are two heads better than one? Pair work in L2 assessment contexts. *Language Testing* 26.3, 325–339.
Toulmin, S. E. (2003). *The uses of argument* (updated edition). Cambridge: Cambridge University Press.
Wagner, E. (2007). Are they watching? An investigation of test taker viewing behavior during an L2 video listening test. *Language Learning and Technology*. http://llt.msu.edu/vol11num1/pdf/wagner.pdf.
Wagner, E. (2010). The effect of the use of video texts on ESL listening test taker performance. *Language Testing* 27.4, 493–514.
Wall, D. & T. Horák (2006). *The impact of changes in the TOEFL examination on teaching and learning in Central and Eastern Europe: Phase 1, the baseline study* (TOEFL Monograph 34). Princeton, NJ: Educational Testing Service.
Wall, D. & T. Horák (2008). *The impact of changes in the TOEFL examination on teaching and learning in Central and Eastern Europe: Phase 2, coping with change* (TOEFL iBT Research Report 05). Princeton, NJ: Educational Testing Service.
Weir, C. (2005a). *Language testing and validation: An evidenced-based approach*. Basingstoke: Palgrave.
Weir, C. (2005b). Limitations of the Common European Framework for developing comparable examinations and tests. *Language Testing* 22.3, 281–300.
Weir, C., B. O'Sullivan, J. Yan & S. Bax (2007). *Does the computer make a difference? The reaction of candidates to a computer-based versus a traditional handwritten form of the IELTS writing component: Effects and impact* (IELTS Research Reports Volume 7, Number 6). Cambridge: Cambridge ESOL.
Wigglesworth, G. & C. Elder (2010). An investigation of the effectiveness and validity of planning time in speaking test tasks. *Language Assessment Quarterly* 7.1, 1–24.
Wise, S. & X. Kong (2005). Response time effort: A new measure of examinee motivation in computer-based tests. *Applied Measurement in Education* 18.2, 163–183.
Wu, J. & R. Wu (2010). Relating the GEPT reading comprehension tests to the CEFR. In W. Martyniuk (ed.), *Aligning tests with the CEFR*. Cambridge: UCLES/Cambridge University Press, 204–224.
Xi, X. (2008). Methods of test validation. In E. Shohamy & N. Hornberger (eds.), *Encyclopedia of language and education* (2nd edn), *Language testing and assessment* (volume 7). New York: Springer, 177–196.
Xi, X. (2010). How do we go about investigating test fairness? *Language Testing* 27.2, 147–170.
Yu, G. (2010). Effects of presentation mode and computer familiarity on summarization of extended texts. *Language Assessment Quarterly* 7.2, 119–136.

STEPHEN STOYNOFF is Professor of English at Minnesota State University, Mankato, in the United States, where he teaches courses in research methods, second language assessment, and second language literacy, in the MA TESL program. He is co-editor, with Carol Chapelle, of *ESOL tests and testing* (TESOL, 2005) and co-editor, with Christine Coombe, Peter Davidson, and Barry O'Sullivan, of *The Cambridge guide to second language assessment* (2012).

https://doi.org/10.1017/S026144481100053X Published online by Cambridge University Press