THE JOURNAL OF ASIA TEFL
Vol. 16, No. 1, Spring 2019, 327-333
http://dx.doi.org/10.18823/asiatefl.2019.16.1.21.327

The Journal of Asia TEFL

http://journal.asiatefl.org/

e-ISSN 2466-1511

An Agenda for Language Assessment Development, Research and Pedagogy

Antony John Kunnan
University of Macau

Introduction

In this inaugural issue of "Assessment Issues," I would like to sketch out a few issues that are not typically dealt with in most assessment-related journals but we would like to deal with them in this *Journal*. First, I want to articulate a research agenda for the 21st century that will serve the interests of testing and assessment (these terms are used interchangeably in this article) agencies, teachers, test takers, test score users, and the public. Second, I want to discuss the issue of institutional responsibility of testing agencies, test developers and test score users through a shift in the curriculum for language assessment students.

A Research Agenda

A comprehensive research agenda for the evaluation of language assessments is necessary for a systematic approach to evaluation.

Assessment Development Process

In Kunnan and Grabowski (2013), we outlined that a typical assessment development and research process is cyclical and iterative process. Figure 1 illustrates the main stages that are involved in the assessment development and research: planning, designing, operationalizing, using and researching. The Figure also shows the iterativity of the stages in the development and research cycle.

During the *planning* stage, the purpose of the assessment is established along with the intended consequences of test use, including the potential decisions that can be made and the impact on test constituents. The *designing* stage includes identifying the TLU domain and outlining the test blueprint and specifications, including any technology considerations. Item- and task-writing, pretesting, and revision then follow in the *operationalizing* stage. Once the test is ready, it can be administered and scored as part of the *using* stage. During this stage, score-based inferences can be made, which translate into decisions about individuals or groups. Once test developers get data from a test administration, research on the items or tasks can be performed during the *researching* stage. Findings from this research can then inform the planning stage in terms of showing support for or evidence against the assessment's

intended purpose or the perceived consequences of test use. Research findings can also inform changes to the designing stage for future administrations of the test. Since this process is not linear, but rather is cyclical and iterative in nature, it is important to note that each ensuing stage in the process can also inform prior stages. For instance, pretesting during the operationalizing stage may uncover issues resulting from the designing stage that need to be addressed before the test can be administered again. Similarly, difficulties revealed during test administration in the using stage may require test developers to modify certain tasks in the operationalizing stage so that listening input is delivered in a more uniform way.

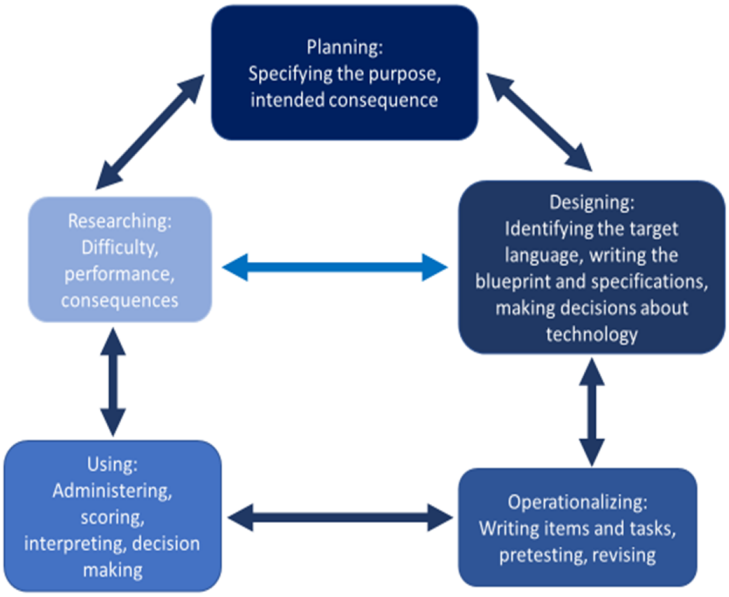

Figure 1. The cyclical nature of assessment development.

Using Figure 1, a framework and agenda for ongoing research can be planned. This should not be an ad hoc plan to satisfy a particular client or customer or satisfy an assessment standard required by funding bodies, Generally, assessment agencies are willing to spend the time and resources to develop and launch an assessment, but many agencies allocate few resources for research and development after the assessment has been launched. Specific plans, procedures, and steps need to be in place for research on validity, reliability, and fairness. In addition, if accommodations (such as extended time) are necessary for test takers with a disability, research in this area has to be conducted in order to arrive at the appropriate type of accommodation.

Assessment development and research staff need to be qualified in language assessment issues, preferably through a university program, and then trained in systematic ways of developing assessments and researching the validity, reliability, and fairness of assessments, and relating all of them to the consequences of assessments. In addition, staff should have the ability to go to public forums (both through academic oral presentations and written journal articles or equivalent) to offer public justification and reasoning of the assessments they have developed and researched. It may be unnecessary to make this recommendation, but given the proliferation of assessments and organizations that are entering this business, it may be salutary.

Assessment score users are members of the larger community who examine scores to make decisions about test takers. Such score users could be school, college, or university teachers who are responsible for placing students into programs, workplace officials who make decisions regarding careers and promotion, and immigration and citizenship officials who decide on applicants' mobility and residence. They need to understand how to read and interpret scores (scores, grades, descriptors, etc.) and understand the limitations of scores (standard errors, cut scores, reliability of scores, etc.). In addition, they need to be

able to translate score reports that are often technical to administrators, parents, and community members in public forums and town hall meetings. On occasion, they should also be prepared to provide depositions in courts related to the strengths of their assessments and contest challenges regarding any weaknesses of their assessments.

Research Focus

Table 1 presents the main areas and sub-areas that can be focused on, with research areas for each. These areas and sub-areas can be articulated as claims and sub-claims that need warrants and backing or rebuttal and backing. The possible list of research areas provides details for the development of specific warrants and backing; it is likely this list may be different based on the claims an assessment agency might make. The order of presentation reflects the linear sequence of assessment development, from receiving the mandate to developing an assessment to the overall evaluation of the assessment. It somewhat distorts the actual process of assessment development in general, but it is used here to show the research areas with more clarity. Figure 1 more accurately places the role of research within the development cycle.

TABLE 1
A research Agenda for Evaluating Language Assessments

Main areas	Sub-areas	Research areas
1. Commission/ mandate	a. Purpose and scope; relationship to instruction; b. Benefit to test takers	Analyses of assessment preparation; opportunity to learn; intended washback; of benefits of the assessment
2. Design	a. Cognitive sources of variance	Analyses of test takers' cognitive process while taking the assessment; irrelevant processes
	b. Content sources of variance	Analyses of specialized knowledge; customs, values, traditions, socialization
	c. Affective sources of variance	Analyses of controversial, upsetting or inflammatory language/content in the assessment
	d. Physical sources of variance	Analyses of audio and visual materials; of needs of test takers with disability
3. Item/Tasks	a. Item/task conceptualization and writing	Analyses of match between specifications and items/tasks
	b. Item/task analyses and reviews, banking and assembly	Analyses of item/task difficulty, discrimination, speedeness; form equivalence
4. Translated/ Adapted forms	a. Comparability of translated adapted or dual-language b. Equating and linking	Analyses of equivalence (construct-bias, method-bias, item-bias); equating and linking
5. Administration	a. Site facilities (room, furniture, equipment, etc.)	Analyses of technology and equipment, computers, and monitors
	b. Site operations (check-in, directions, monitoring test takers, check-out, storage or return of materials)	Analyses of administration staff; procedures for check-in (biometric data, assistive technologies, hearing aids, etc.); of security and fraud check; procedures for dealing with irregularities
	c. Test takers with disabilities	Analyses of accommodations: Extended time, frequent breaks, large font size, etc.
	d. L2 test takers taking math and science assessments in L1	Analyses of accommodations: from glossaries to dual language presentations
6. Scoring, interpretation and decision-making	a. Human and automated scoring	Analyses of consistency among raters; of rating systems; comparability of feedback; differential item/task functioning; standard setting
	b. Washback and instruction	Analyses of intended and unintended washback of the assessment on teaching and learning
	c. Consequences	Analyses of standard setting; pass-fail rates; differential effects on test taker groups; public policy
7. Overall evaluation	a. Benefit to immediate stake-holders b. Benefit to the community	Analyses of pros and cons of the assessment to the community

Refocusing the Curriculum

The development and research of assessments, whether they are done by individual teachers or a committee of teachers, needs careful thinking in terms of assessment conceptualization, blueprint design, item/task writing, and research components. Different levels of expertise may be required based on whether the assessment that is being developed as a large-scale assessment with high or low stakes. Brown and Bailey (2008) documented the state of language testing courses in 2007 through a questionnaire study. They reported on course characteristics including topics (such as test consistency and validity), statistical concepts (such as mean, median, mode, standard deviation, variance), and required books. The study showed that most of the courses focus on traditional components and offer few newer areas of interest including alternative assessment. The conclusions drawn are from training curricula for Diploma, B.A., M.A. programs and research curricula for Ph.D. programs. In thinking further about the components, Kunnan (2017) argued that optimum assessment knowledge and techniques needed by new students to the field fall into three approaches with some overlap between them.

Approach 1: The Traditional-Engineering Curriculum

This approach includes:

(1) An overview of *skills and components* to be assessed (listening, speaking, reading writing; pronunciation, grammar, vocabulary; pragmatics, integrated skills, etc.)
(2) *Techniques for test development*, creation and scoring (item writing and revision, using different response formats, writing rubrics for scoring, designing score reporting formats, etc.)
(3) *Skills for improving, editing and assembling* of items and tests (planning revision cycles, writing protocols for item banking, etc.)
(4) *Research themes*: reliability (internal consistency, inter-rater), content, construct and predictive validity, item and test bias (Differential Item/Test Functioning analysis)
(5) *Research methods*: Quantitative (classical true score theory, item response theory) and qualitative (introspective analysis, conversational analysis) approaches for research (procedures for item and test difficulty, discrimination, bias; and analyzing cognitive and affective processes in test taking)
(6) *Training with software*: SPSS, ITEMAN, SAS for descriptive and inferential statistics, FACETS, WINSTEPS for analyzing ratings, MULTILOG for bias analysis, etc.
(7) *Technical and score user manuals* (writing of a parts of a technical manual that includes a description of the test, administrative, and scoring and reporting details, accommodations for test takers with disability, a report of research studies that support the claims of the test, and samples of tests, scoring rubrics, and score reports; and pricing of different services – standard prices, expeditated score-reporting, human scoring, detailed score report and diagnostic feedback, etc.).

Approach 2: The Innovative-Design Curriculum

This approach includes:

(1) *Innovative design* in items and tests, scoring and reporting (integrated tasks such as listening-writing, reading-speaking, tasks that match professional work such as assessing the English language ability of aviation professionals, etc.)
(2) *Use of new technologies* (tasks that involve audio-video, multi-media, multi-modal, automated scoring of writing and speaking, tests on tablets, etc.)
(3) *New needs and uses* (immigration, citizenship, asylum, and forensic purposes, etc.)

(4) *Research themes*: authenticity, instructiveness; cognitive diagnostic feedback, dynamic assessment, etc.
(5) *Research methods*: new analyses using structural equation modeling, hierarchical linear modeling, multi-level modeling, discourse and conversational analyses, etc.
(6) *Training with advanced software*: AMOS, EQS, Mplus for modeling, etc.

Approach 3: The Ethical-Critique Curriculum

This approach ought to include:

(1) *Understanding of the history* of language assessment (from the Chinese civil service examinations, *le baccalaureate*, *abitur* to popular modern language assessments such as the TOEFL, iBT, FCE, CAE, CPE, IELTS, SAT, etc.)
(2) *Understanding the political motivations and legal bases* for assessments (the U.S. Naturalization test, similar tests in the U.K, Australia, Germany, South Korea, etc.)
(3) *Understanding the social and cultural assumptions* of assessments (knowing the semiotic domain, and having embodied experiences)
(4) *Understanding the philosophical underpinnings* of assessments (utilitarian, social contract, deontological, pragmatist and humanist approaches)
(5) *Working with hypothetical scenarios* or case studies of assessments that need judgments using moral or ethical philosophy.
(6) *Using ethical principles* to evaluate assessments (parochial versus global ethics, etc.)
(7) *Writing defenses and critiques* of assessments and assessment practice (assessment reviews).

It may be obvious to a language assessment professional that the focus of most training programs (Certificate, B.A., M.A., or Ph.D.) have been on the first approach although a few programs address some aspects of the second approach. Aspects of the third approach are generally ignored or rarely offered. This could be due to a number of reasons: the first approach lays the foundation of training and therefore has to be included in all programs; the focus of particular programs could be based on the expertise of the faculty of the program (which is generally in the first strand); the second approach of innovation in design, tasks, and research is based on new purposes, contexts, and technologies; the third approach is the newest and requires the most time to develop as it is interdisciplinary with subjects in humanities and social sciences. These strengths and weaknesses of language assessment curricula are also reflected in published works (in journals and books) and in conference themes and presentations. In addition, most programs of study only have one language assessment course. In these contexts, Approach 1 is the most likely choice with a few aspects of Approach 2 thrown in, if there is space created in the curriculum for these aspects. Ideally, of course, at least three or more courses of study would be recommended if students are training to become professional assessment developers and/or researchers.

Fulcher (2012) offered an expanded working definition of language assessment literacy which has been in the forefront of discussions on the topic:

> The knowledge, skills and abilities required to design, develop, maintain or evaluate, large-scale standardized and/or classroom-based tests, familiarity with test processes, and awareness of principles and concepts that guide and underpin practice, including ethics and codes of practice. The ability to place knowledge, skills, processes, principles and concepts within wider historical, social, political and philosophical frameworks in order understand why practices have arisen as they have, and to evaluate the role and impact of testing on society, institutions, and individuals.

Fulcher (2012) showed the different layers of knowledge work together: knowledge of skills and abilities (termed Practices), processes, principles and concepts (termed Principles), and historical, social,

political, and philosophical frameworks (termed Contexts). While this framework does not completely overlap with the curricula presented earlier, it is clear that the Traditional-Engineering curriculum can be matched with the Practices, and the Ethical-Critique curriculum can be matched with the Contexts.

The challenge for language assessment training programs then would be to offer a comprehensive approach incorporating the most important components of the three curricula or Fulcher's layers based on needs and orientation of the programs. It is hoped however that all programs will include elements of the ethical-critique curricula or historical, social, political and philosophical context so that students and professionals are aware of the context of their work as well as their responsibilities (like healthcare professionals like medical doctors, pharmacists, hospital nurses, engineers, and computer scientists are increasingly asked to).

New Topics

Based on these discussions, new topics that would be welcome for publication in these pages include:

- Integrated skills tasks such as listening-writing, reading-speaking, tasks that match professional work such as assessing language proficiencies of aviation professionals, health professionals, culinary professionals, etc.
- New technologies in tasks that involve audio-video, multi-media, multi-modal, etc.
- Automated scoring of writing and speaking, relating how automated scoring can be combined with human scoring (see Hoang & Kunnan, 2016; Liu & Kunnan, 2016)
- New needs and uses of assessments in areas such as immigration, citizenship, asylum, and forensic purposes, etc. (see Kunnan, 2009a, 2009b; Kunnan, 2012; Kunnan, 2017, Chapter 8)
- Philosophical underpinnings of assessments (utilitarian, social contract/deontological, pragmatist and humanist approaches) (see Kunnan, 2017, Chapters 3 & 10)
- Political motivations, social and cultural assumptions and legal bases for assessments (see Kunnan, 2017)
- Ethical principles to evaluate assessments from a particular or global perspective (see Kunnan, 2017)
- Court challenges and critical reviews of assessments and assessment practice (see Wagner & Kunnan, 2015).

Conclusion

I am hoping that these ideas regarding assessment development, research and pedagogy would propel some new thinking in this field and can result in publication of issues related to language assessment in these pages in the future.

The Author

Antony John Kunnan is Professor of Applied Linguistics and Associate Dean of the Faculty of Arts and Humanities at the University of Macau. His area of specialization is language assessment and he has give n talks at over 130 conferences in 35 countries. His recent books are: *The Companion to Language Assess ment* (4-volume edited set of 140 papers; Wiley, 2014) and *Evaluating Language Assessment* (Routledge, 2017). He is the past president of the International Language Testing Association, the founding president of the Asian Association for Language Assessment, and the founding editor of *Language Assessment Quarterly*.

E-mail: akunnan@umac.mo

References

Fulcher, G. (2012). Assessment literacy for the language classroom. *Language Assessment Quarterly, 9,* 113-132.

Hoang, G., & Kunnan, A. J. (2016). Automated essay evaluation for English language learners: A case study of my access. *Language Assessment Quarterly, 13*, 359-376.

Kunnan, A. J. (2009a). Politics and legislation in citizenship testing in the U.S. *Annual Review of Applied Linguistics, 29,* 37-48.

Kunnan, A. J. (2009b). The U.S. naturalization test. *Language Assessment Quarterly, 6,* 89-97.

Kunnan, A. J. (2012). Language assessment for immigration and citizenship. In G. Fulcher & F. Davidson (Eds.), *The handbook of language testing* (pp. 152-166). New York, NY: Routledge.

Kunnan, A. J. (2017). *Evaluating language assessments*. New York, NY: Routledge.

Kunnan, A. J., & Grabowski, K. (2013). Large scale second language assessment. In M. Celce-Murcia, D. M. Brinton, & M. A. Snow (Eds.), *Teaching English as a second or foreign language* (4th ed., pp. 304-319). New York, NY: Heinle Cengage.

Liu, S., & Kunnan, A. J. (2016). Investigating the application of automated writing evaluation to Chinese undergraduate English majors: A study of *WriteToLearn. CALICO Journal, 33*, 71-91.

Wagner, E., & Kunnan, A. J. (2015). Duolingo English test. *Language Assessment Quarterly, 12*, 320-331.