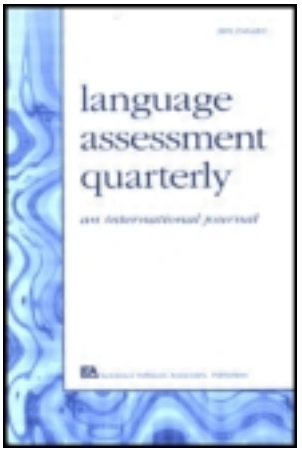

Language Assessment Quarterly
Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/hlaq20

Assessment Literacy for the Language Classroom

Glenn Fulcher [a]
[a] University of Leicester

Version of record first published: 07 May 2012.

To cite this article: Glenn Fulcher (2012): Assessment Literacy for the Language Classroom, Language Assessment Quarterly, 9:2, 113-132

To link to this article: http://dx.doi.org/10.1080/15434303.2011.642041

Language Assessment Quarterly, 9: 113–132, 2012

ISSN: 1543-4303 print / 1543-4311 online
DOI: 10.1080/15434303.2011.642041

Routledge
Taylor & Francis Group

ARTICLES

Assessment Literacy for the Language Classroom

Glenn Fulcher
University of Leicester

Language testing has seen unprecedented expansion during the first part of the 21st century. As a result there is an increasing need for the language testing profession to consider more precisely what it means by "assessment literacy" and to articulate its role in the creation of new pedagogic materials and programs in language testing and assessment to meet the changing needs of teachers and other stakeholders for a new age. This article describes a research project in which a survey instrument was developed, piloted, and delivered on the Internet to elicit the assessment training needs of language teachers. The results were used to inform the design of new teaching materials and the further development of online resources that could be used to support program delivery. The project makes two significant contributions. First, it provides new empirically derived content for the concept of assessment literacy within which to frame materials development and teaching. Second, it uncovered methodological problems with existing survey techniques that may have impacted upon earlier studies, and solutions to these problems are suggested.

CHANGING TIMES

The first decade of the 21st century has seen a phenomenal increase in the testing and assessment responsibilities placed upon language teachers. There are arguably three primary reasons for this, two of which are external to the field and one of which is internal.

First is the increased use of tests and assessments, both externally mandated and locally developed, for the purposes of accountability. Malone (2008) identified the No Child Left Behind legislation in the United States, and the widespread implementation of the Common European Framework of Reference in Europe, as major change factors. Similarly, McNamara and Roever (2006) referred to the Common European Framework of Reference as "the most comprehensive example of policy-driven assessment yet seen" (p. 212). Although policymakers have always

Correspondence should be sent to Glenn Fulcher, University of Leicester, School of Education, 162 - 164 Upper New Walk, Leicester, LE1 7RF United Kingdom. E-mail: gf39@le.ac.uk

sought to control educational practice through the use of tests, we are more aware than ever of the power of tests (Shohamy, 2001), and the way in which they are used in political systems to manipulate the behaviour of teachers and hold them accountable for much wider policy goals. Brindley (2008) noted that it is the externally mandated nature of tests that make them such attractive political tools. Policymakers frequently believe that changes can be implemented relatively quickly and cheaply without having to undertake curriculum development or change classroom practices through teacher education programs. Often motivated by fears for the economic future of the country without an appropriately skilled workforce, tests are perceived to address the need to raise educational standards by introducing transparent means of accountability (Fulcher, 2009). Tests of international literacy, such as the PISA program, are also used in an accountability role (McGaw, 2008). The results of such tests not infrequently lead to national educational reform, as was the case in Germany following the 2000 tests (Fertig, 2003).

Far from being immune to the use of tests in this way, teachers are the target of the intended effects. Yet teachers often seem unable to affect the policy, change the intended effect, or resist external impositions when they are regressive. This may in part be because of a lack of conceptual assessment tools to evaluate and construct counter arguments, or the practical skills to investigate tests, test use, and deal with intended changes.

The second major reason is the rapid expansion of the use of language tests as part of national immigration policy, as surrogates for immigration policies, or components of citizenship tests (Kunnan, 2009; McNamara, 2008). Globalization and an increase in international migration have led many countries to become concerned with perceived threats to national identity, which in many cases is closely linked with language in the minds of policymakers (Extra, Spotti, & Van Avermaet, 2009; Hogan-Brun, Mar-Lolinero, & Stevenson, 2009: Slade & Möllering, 2010). Teachers are not the intended effect of this use of tests, but it impacts upon the nature of their work, particularly with regard to the demand for test preparation classes and the expectation on the part of the "client" that they will get the results they require for international mobility. A teaching environment is created that is driven by the economic value placed on test scores. The growing interest in the washback of tests on what teachers do has also starkly illustrated the need for teachers to be aware of how their work is often shaped by testing policies and practices (Cheng, 2008; Wall, 2005, 2012).

The third reason is internal to the field. It has long been argued that assessment for learning in some guise is an essential component of classroom practice (Black & Wiliam, 1998). Recently this has led to an increased focus on assessment in language programs and its role in enhancing learning (Rea-Dickins, 2006, 2008). Most frequently this is framed in terms of Vygotsky's notion of the zone of proximal development in which noticing the gap between what a learner can do now, and the target production, is the key to making progress. Noticing the gap may be as a result of focus on form approaches (Long, 1991), or much more explicit interventionist approaches like Dynamic Assessment (Lantolf, 2009; Lantolf & Poehner, 2008). The larger agenda is to increase learner motivation through the establishment of a culture of success, promoted by organizations such as the Assessment Reform Group (http://www.aaia.org.uk/afl/assessment-reform-group/) through documents like *Assessment for Learning: 10 Principles* (Assessment Reform Group, 2002). Language teachers will in future be expected to have a range of strategies at their disposal to implement classroom assessment, and evaluate its success.

If language teachers are to understand the forces that impact upon the institutions for which they work and their daily teaching practices, and to have a measure of control over the effects

that these have, it is important for them to develop their assessment literacy. Precisely what knowledge, skills, and principles are components of this literacy is still under discussion.

DEFINITIONS AND DEBATES

The term "assessment literacy" (Stiggins, 1991, 1997) has become accepted to refer to the range of skills and knowledge that stakeholders need in order to deal with the new world of assessment into which we have been thrust. Yet there is little agreement on what "assessment literacy" might comprise, despite an increasing diversity of approaches that are recommended to encourage its development (e.g., Walters, 2010). This article describes a research project designed to collect empirical data from language teachers, rather than teachers of language testing, to discover what learning needs they have in language testing and assessment. The intention was to use the outcome of the needs analysis to give empirically derived content to the concept of assessment literacy and inform the design and construction of new materials that can be used in language testing education programs.

The range and number of stakeholders who require a level of assessment literacy has grown. Taylor (2009) included university admissions officers, policymakers and government departments, in addition to those professionally engaged with testing and assessment, and teachers. Yet there are few textbooks and learning materials available for nonspecialists or those new to testing and assessment. Echoing Brindley (2001, p. 127), Taylor argued that most available textbooks are "highly technical or too specialized for language educators seeking to understand basic principles and practice in assessment" (p. 23). It also seems that teachers of language testing are slow to abandon texts with which they are familiar. Studies by Bailey and Brown (1996) and J. D. Brown and Bailey (2008) reveal that textbooks are rarely changed. Similarly, the content of language testing courses does not appear to change very much, perhaps reflecting the same conservatism. J. D. Brown and Bailey showed that emphasis continues to be placed upon the same topics: critiquing and analysing tests, measuring the four skills, validity (particularly a more traditional "types" approach), item analysis (facility value, discrimination index, and content analysis), and basic test statistics including descriptive statistics, reliability, and error.

These topics would fit well into what Davies (2008) would call a "skills + knowledge" approach to assessment literacy. "Skills" refers to the practical know-how in test analysis and construction, and "knowledge" to the "relevant background in measurement and language description" (Davies, 2008, p. 328). There is evidence that this model is not only widely followed in North America, Europe, and Australia but also has been adopted by other countries that have industrial testing operations, such as China (Jin, 2010).

Davies (2008) argued that what is missing is a focus upon "principles," or the reasons for testing or assessing, explored within a social and historical context. This would also include issues such as the ethics of testing, test fairness, and the role of tests in political decision making in controversial areas like immigration. The historical context is also important. Young disciplines like language testing are constantly reinventing the wheel, and as Spolsky (1995) noted, "our field has been remarkably ahistorical" (p. 150).

The earliest attempt to define assessment literacy for teachers was produced by the American Federation of Teachers (1990), although the term "assessment literacy" was not in use at the time. The competencies included selecting assessments, developing assessments for the classroom,

administering and scoring tests, using scores to aid instructional decisions, communicating results to stakeholders, and being aware of inappropriate and unethical uses of tests. It can be seen that Davies' notion of "principles" was present in this early document, although there is little evidence of it having impacted upon the teaching of language testing, either in the textbooks of the time or the courses as surveyed by Bailey and Brown. Brindley (2001) was the first language tester to visit the topic of assessment literacy. He argued for a focus on "curriculum-related assessment." Even when textbooks discuss the needs of classroom teachers, they frequently describe techniques that are drawn from large-scale standardized testing, many of which are not applicable to the classroom. For example, the kinds of statistical analysis presented in the most popular text (Hughes, 1989/2003) require large sample sizes, to which teachers rarely have access. The assumption underlying the selection of material is that teachers will primarily be interested in the construction and evaluation of norm-referenced tests. Although it seems reasonable that all teachers should have a grasp of the workings of large-scale testing and the test development practices associated with it, there does appear to be a prima facie case to introduce more of a balance between normative and classroom-based approaches.

Brindley also argued for a discussion of the social context of testing, as well as the requirements to define and assess proficiency, construct and evaluate tests, and use tests and assessments to measure attainment against a curriculum. Perhaps most important, Brindley recognized that teachers work within time and resource constraints and urged testing educators to recognize that they must develop flexible approaches to their assessment practices. These recommendations prefigure Davies' concern for an expansion of the traditional content of books on language testing to meet the emerging needs of the 21st century.

Somewhat more radical is the analysis offered by Inbar-Lourie (2008a). Rather than adding a concern for social context, she placed social context at the heart of assessment and assessment literacy. She argued that there is a great divide between the cultures of "testing" and of "assessment." The former is the field of psychometrics with its positivist view of the world; the latter is part of a learning culture, and "learning cultures are grounded in interpretive epistemology which views reality as the subject of social construction" (Inbar-Lourie, 2008a, p. 387). The argument is that a new "assessment culture" is required, where people "share epistemological suppositions about the dynamic nature of knowledge" (p. 387). This would foreground formative assessment practices, the study of the impact of different assessment methods, and training for teachers to become facilitators of learning through assessment. The view that the "testing" and the "assessment" cultures are incompatible is similar to that of Shepard (2000, p. 8), who compared them to the dark and the good side of the force, with traditional language testers cast in the role of Imperial storm troopers. The problem with postmodern approaches of this kind is that they end up being more coercive than the forces they intend to replace. For example, Inbar-Lourie (2008b) stated, "Assessment culture can comprise part of the culture of educational organizations such as schools, providing that the members of the organization adopt the beliefs and assumptions regarding the nature of assessment and its role in the learning process" (p. 288). According to this view, acquiring assessment literacy therefore means teachers must undergo a "profound perception change" (Inbar-Lourie, 2008b, p. 293) as they accept that all knowledge and meaning is socially constructed.

Although it is arguably the case that language testing and assessment is a practice that has evolved to solve social problems and that the social and consequential aspects of assessment are very important, we should not be tempted to adopt a standpoint epistemology that reduces

meaning to individual and group perception. Those who are tempted should recall Nietzsche's (2005) *Twilight of the Idols,* in which he provides a thumbnail sketch of the history of western philosophy as a story of how the "true world" became a fable: "The true world is gone: which world is left? The illusory one, perhaps? . . . But no! *We got rid of the illusory world along with the true one*!" (p. 171). In short, if reality or knowledge is what we perceive and may temporarily construct with others, there can be no reality or knowledge at all, and no criteria by which to evaluate the appropriacy of competing validity (or ethical) arguments. This postmodern approach to assessment literacy represents a swing of the pendulum to another extreme that will ultimately fail teachers as profoundly as an approach that prioritizes the large-scale standardized psychometrically driven paradigm.

RESEARCH INTO THE PEDAGOGY OF LANGUAGE TESTING

Research into assessment literacy is in its infancy. Bailey and Brown (1996) and J. D. Brown and Bailey (2008) looked at the content of language testing programs and the text books used, discovering that little had changed over the decade of the research. This research focused entirely on teachers of language testing reporting on the content of their courses. Plake and Impara (1993, 1996) reported on a survey of assessment literacy in the United States, designed to measure teachers' knowledge of the components of the American Federation of Teachers (1990) *Standards*. The constructs included the following:

1. Choosing an assessment method appropriate for instructional decisions.
2. Developing assessment methods appropriate for instructional decisions.
3. Using assessment results when making decisions about individual students, planning instruction, developing curriculum, and improving schools.
4. Developing valid pupil-grading procedures.
5. Communicating assessment results for students, parents, other lay audiences, and other educators.
6. Recognizing unethical, illegal, and other inappropriate methods and uses of assessment information. (Plake & Impara, 1996, p. 54)

Using a 35-item test, the researchers discovered that on average teachers were responding correctly to just 23.2 items, which they argue shows a low level of literacy. There was a correlation between experience and score, but it was not possible to detect the assessment literacy needs of teachers because the item facility indexes and point biserial correlations were erratic, showing little consistency of items within intended constructs. Although the authors did not conduct exploratory factor analysis, from the available data it seems likely that the constructs would have little operational structural integrity.

Hasselgreen, Carlsen, and Helness (2004) and Huhta, Hirvalä, and Banerjee (2005) conducted a survey designed to uncover the assessment training needs of teachers in Europe. Although there were problems separating "teachers" from "trainers" and "experts" in the data, and some countries were more heavily represented in the data than others (e.g., Finland), the research seems to have uncovered the following needs: portfolio assessment, preparing classroom tests, peer- and self-assessment, interpreting test results, continuous assessment, giving feedback on work, validity, reliability, statistics, item writing and item statistics, interviewing, and rating.

The studies undertaken to date have been useful within the parameters that were set. However, all of these studies suffer from their utilization of primarily closed-response items, which lend themselves only to quantitative treatment and which tend to produce similar responses from all participants. Specifically, respondents tend to say that everything presented to them is important, resulting in little variation. This effect may be partly due to the predisposition to give answers that are likely to be pleasing to the researcher and/or to the nature of the sample, especially where random sampling is not feasible. This study has attempted to gain a general view of the assessment needs of language teachers that can be used as a basis for the development of new educational materials through both closed- and constructed-response items and used a number of innovative design features that encourage teachers to express needs independently of the predigested response options. As such, it has attempted to fill a gap in the field, and provide further substantive definition to the construct of "assessment literacy."

METHODOLOGY

Funded by the Leverhulme Trust, a survey instrument was developed (see the appendix) to address the research question, "What are the assessment training needs of language teachers?" The explicit purpose of the study was to inform the development and writing of a new text on language testing for teachers, along with associated materials that would be made available on the Internet. The survey was piloted using 24 international language teachers. The pilot study identified problems with the wording of a number of items, which were then modified prior to the main study. In particular, it was observed that respondents relied far too heavily upon the concepts and wording of the closed-response items. In short, if closed-response items are presented to respondents before they are offered the opportunity to formulate their own ideas and create a personal frame of reference, the language of the closed-response items is adopted and given back to the researcher in subsequent constructed-response items. As a result of this finding, one set of constructed-response questions (Questions 2 and 3) was placed before the closed-response items. The intention was to get teachers to think about their assessment needs in their own words before being presented with the language of the closed-response items that is given by the survey designer. On a second pilot it was discovered that reliance on the wording and concepts presented in the closed-response items decreased throughout the survey. A second design innovation was to structure the constructed-response questions in pairs. The first of each pair was intended to get teachers to reflect on their experiences (e.g., in training, or reading textbooks), whereas the second asked them to think about what they still need or require (to do their job, or would like to see in educational materials). It was observed that the language used in responding to the second of the two questions was also far less likely to echo the wording or terminology of the closed-response items. These disadvantages of closed-response items need to be taken into account when using surveys in similar needs analysis studies, and could usefully be added to warnings in the research methods literature (J. D. Brown, 2001, p. 37).

The questionnaire was delivered over the Internet using Lime Survey Software (http://www.limesurvey.org/) and was widely advertised through professional organizations and discussion lists. The intended population was described as "language teachers," but the sample was essentially self-selecting. This has both positive and negative implications. On one hand, each respondent had an interest in language testing, was motivated to answer the questions, and had something to say. On the other hand, any self-selecting sample is bound to be biased in some way,

perhaps most obviously in terms of their interest in the topic. Yet conducting a bias analysis is not possible because there is no way to describe "nonrespondents" in an online environment of this kind. However, to encourage some to respond who may not otherwise have done, each respondent was entered into a prize-draw. As respondents completed the questionnaire, their responses were automatically loaded into a database on the web server, from which they could be downloaded for use in Microsoft Excel or SPSS. Answers to the constructed-response questions were saved for analysis as PDFs.

Data were collected between June and September 2009, and in total there were 278 responses. Sixty-nine percent of respondents were female, and most fell in the 41 to 45 age range; however, 29% of all respondents were in the 30 to 40 age range, and 17% were in their 20s. Eight percent of the sample held a relevant BA degree as their highest qualification, 47% held an MA degree, and 38% a doctorate. The remainder had high-school-level qualifications. The fact that 85% of respondents hold a higher degree may not be very surprising, as an MA degree is rapidly becoming a required qualification for language teachers who seek career progression; however, it does indicate that the respondents to the survey were largely from the more highly educated subpopulation. Of those who held a doctorate, most taught languages at a university. Respondents were from Australia and New Zealand (13.5%), North America (13.5%), South America (5.4%), the Middle East (2.7%), the Far East (16.2%), and Europe (37.8%). The range of countries represented suggests that the sample is genuinely international. The number of languages spoken as a first language (L1) reflected the geographic spread of the sample, although there were a number of English L1 speakers in the sample working in other countries. Only 12% of the sample reported speaking only one language—invariably English, whereas 48% reported being competent in two or more languages in addition to their L1. Respondents were also asked to rate their knowledge and understanding of language testing (Question 9). The mean rating was 3.65 with a standard deviation of .9 (on a scale of 5). This outcome is not unexpected for a self-nominating sample, which is likely to have an interest in language testing in order to voluntarily complete the questionnaire.

In the analysis of both the quantitative and qualitative responses, data reduction strategies were employed to arrive at meaningful interpretations. Closed-response items on the survey were analysed using exploratory factor analysis to identify components of assessment literacy. Once factors were extracted descriptive statistics and reliability were calculated for each, and an attempt to arrive at meaningful factor labelling undertaken.

Analysis of qualitative data is always more problematic, as there is no canonical approach. This research followed the procedures recommended by Miles and Huberman (1994, pp. 58–69). The first step was to construct a data-coding matrix using the four factors identified from the factor analysis at the highest level and the questions within each factor as a finer coding. The entire coding matrix is self-evident from this example for the first factor:

A. Test Design and Development
   1. Writing items and tasks
   2. Writing test specifications
   3. Developing rating scales
   4. How to decide what to test

and so on, following the factor structure in Table 1. This a priori coding system was used in a first pass over the constructed response data in an attempt to identify the range and number of responses that mentioned as important the features included in the closed-response items.

TABLE 1
Factor Analysis of the Survey

		Factor				
Code	*Name*	*1*	*2*	*3*	*4*	h^2
E	itemstasks	.891	.137	.224	.124	.877
D	specs	.695	.132	.204	.058	.541
M	ratingscales	.610	.318	.280	.224	.601
C	what2test	.596	.157	.453	.233	.639
B	design	.573	.186	.294	.215	.522
F	evaluating	.482	.338	.318	.321	.550
L	stats	.193	.715	−.012	.308	.643
Q	standard	.138	.622	.277	.087	.483
W	edmeas	.212	.615	.237	.001	.453
P	largescale	.302	.610	.100	.159	.498
H	analysis	.424	.597	.147	.159	.583
G	scores	.467	.562	.205	.266	.646
V	uses	.093	.445	.400	−.023	.358
N	scoreclosed	.375	.400	.145	.028	.322
A	history	−.141	.280	.124	.121	.128
S	washback	.243	.178	.573	.065	.149
R	prep	.145	.084	.556	.023	.330
O	classroom	.208	.019	.537	.164	.358
U	ethics	.218	.454	.536	.146	.561
T	admin	.098	.365	.532	.061	.416
I	selecting	.344	.266	.458	.107	.410
K	validation	.536	.341	.198	.710	.946
J	reliability	.456	.397	.237	.647	.841
	Eigenvalues	4.119	3.791	2.766	1.502	12.178

However, it was anticipated from the pilot and the redesign of the survey instrument that participants would raise additional features. As such, it was necessary to add inductive coding categories that appeared salient to the researcher. The additional inductive categories produced for the analysis were (a) conceptual explanation, (b) language testing as process, and (c) social impact. The constructed response items were then read again, and discourse chunks that were relevant to these interpretations coded. Particularly clear examples of each were noted for illustration of the category.

FINDINGS AND DISCUSSION

Closed-Response Items

Question 4, with its 23 subquestions, generated the only data that can be analysed statistically. Cronbach's alpha was .93, indicating a high level of reliability. As the Kaiser–Meyer–Olkin measure of sampling adequacy was also high (.89), a maximum likelihood exploratory factor

analysis with varimax rotation was conducted in order to further explore the responses, following J. D. Brown (2001, pp. 184–187; 2010). The results are presented in Table 1, where the item labels in the left hand column relate to the subquestions (A–W) in Question 4 (see the appendix). The table shows that four factors with an eigenvalue greater than 1 emerge, accounting for 52.95% of the reliable variance, with totals of 17.91% for Factor 1, 16.48% for Factor 2, 12.03% for Factor 3, and 6.53% for Factor 4.

Taking an arbitrary (but common) cutoff at .4 as a significant loading (see Child, 1990), although there are some areas of inevitable overlap between factors, it would initially appear reasonable to label the first factor as "test design and development," to which notions of reliability and validity are certainly related, rather than thinking of them being purely a post hoc activity. The second factor is usefully interpreted as "large-scale standardized testing" including the use of scores from such tests, the third factor as "classroom testing and washback," and the final factor as "validity and reliability." It is interesting that ethics loads on both the standardized testing and classroom testing factors, perhaps indicating participants' perception that principles are important to all testing practice. Nevertheless, as McNemar (1951) would remind us, the interpretation of such factor loadings is more of an art, if not wishful thinking, than a science. Treated cautiously, however, these four hypothetical factors may indicate broad subject areas that should be covered by educational materials for students of language testing. Each category would not only include the knowledge and skills required to undertake language testing work but also would cover the principles and history that free practitioners to make informed, ethical decisions. The reliability and descriptive statistics for the four factors are presented in Table 2.

Like the European study (Hasselgreen et al., 2004), this study suffered from a lack of differentiation between respondents, such that even with a reasonable sample size it proved impossible to achieve enough power to discover if certain subgroups had specific needs. Furthermore, it appears to be the case on most published quantitative surveys that most respondents feel that "everything is important," which makes further analysis difficult. Nevertheless, a small number of relevant observations may be made from this data.

With regard to test design and development, there was significant variation depending upon responses to the question asking teachers to estimate their level of knowledge of language testing (Question 9). Respondents who said that their knowledge was poor tended to rate their need for training in test design and development slightly lower (.52 on average; $p = .03$) than those in other categories. However, the practical significance of this finding is low ($\eta^2 = .096$), despite power achieving .73. The finding is also not intuitively satisfying from a pedagogical perspective, unless teachers in this category see their role primarily as selecting tests for use with students or preparing them to take tests, rather than designing and developing their own tests. Even if this

TABLE 2
Reliability and Descriptives for the Four Factors

Factor	*Cronbach's α*	*M*	*SD*	*SE*
Test design and development	.89	4.44	.68	.05
Large-scale standardized testing	.86	3.88	.69	.05
Classroom testing and washback	.79	3.82	.73	.06
Validity and reliability	.94	4.49	.85	.07

were to be the case, there could be little justification for ignoring test design and development in pedagogically oriented materials. Second, a significantly higher response to items related to the large-scale standardized testing factor was found for the 14 teachers who reported that they were currently employed by examination agencies (.4 on average; $p = .03$), although once again the practical significance was low ($\eta^2 = .052$), and power was moderate (.64). Nevertheless, this small effect does make pedagogic sense and needs to be taken into account for this group of potential users of materials. Finally, with regard to the study of validation and reliability, the age of respondents was a factor with those older than age 60 ($n = 9$) reporting a significantly lower need to know more about related topics (.5 on average; $p = .03$). However, practical significance was again low ($\eta^2 = .096$) despite adequate power (.75).

If quantitative studies of this nature are to be used in the future, it is clear that much larger sample sizes will be required to generate the power necessary to discover any differences in response between subsections of the test-taking population. Even then the tendency to think that everything is important is likely to result in similarly low effect sizes. Indeed this may be unavoidable in any study where it is not possible to have genuinely random samples from a clearly defined population of teachers. Where random sampling is not possible it seems that much more can be learned from constructed-response questionnaire items that attempt to explore the assessment needs of respondents in ways that do not require the use of predigested lists. As discussed in the Methodology section, this survey used a larger number of open-ended data collection techniques that were deliberately designed to provide opportunities for free expression.

Constructed Response Items

Questions 2 and 3 were asked before the respondents were exposed to the closed-response options, and so were likely to generate the most open, unaffected responses from teachers. The two questions were designed to get teachers to compare their experience of learning about language testing with their perception of what they still need to effectively use assessment in their current posts. Responses to Question 2, which asked when they had last studied language testing and what they had covered, reflected the flip side of the findings of J. D. Brown and Bailey (2008): Irrespective of where they are in the world, teachers report having studied the same topics that teachers report teaching, with a particular emphasis upon critiquing language tests. All responses to this question could be understood in terms of a priori coding, with more than 80% of respondents' comments being coded as "B. Large scale standardized language testing." At the next level of granularity category "5: test analysis" was listed by everyone. However, the responses to Question 3, which asked what skills the teachers felt they still needed, revealed a different perspective. The most frequently mentioned topic was that of statistics. The emphasis, however, was not on simple calculation of basic test statistics but on developing a conceptual understanding of the statistics: Why are we doing this? These comments were coded as "conceptual explanation" and occurred in 35% of responses. A typical illustrative comment from the data was "I don't understand statistics, but I know they can be useful. I need it explaining conceptually, rather than just calculations." This was part of a trend seen across the data that implies the teaching of statistics for assessment needs to be embedded within a larger narrative that relates them to their historical context, and a philosophy of language and measurement (see Fulcher, in press).

The other areas of perceived need were much more practical, and for the most part could be coded within the a priori system. Of most concern was the ability of teachers to "check

reliability and validity of tests at each stage in development." However, in cases like this there was a double coding, using the inductive category of "language testing as process" because of the clear awareness that there are "stages of development." This feature was noted in 26% of responses, and it may be interpreted to suggest that pedagogically oriented materials take teachers through a process of development and implementation from beginning to end. Indeed, one respondent saw reliability and validity as concerns from test specifications to operationalization: "Issues to do with reliability and validity in language testing; the test writing process from the creation of the test specifications through to the trialling and administration and marking of tests." Other comments on language testing as "process" suggested that discussion of validation practices might extend to aligning tests to the curriculum or to external standards, as well as more practical matters like training raters and ensuring that tests are delivered in a "fair" way.

Some 21% of teachers also expressed the concern seen in Brindley and Taylor with regard to balance between classroom and standardized testing, which was coded as "C. Classroom Assessment 3. Classroom focus." One participant wrote that we need "differentiation between classroom assessments, formative assessment, and large scale assessment when discussing key issues." Reliability and validity was seen as relevant to each context, but the respondents thought that they should be treated differently in each context. This is a sophisticated insight that has been treated to some degree in the technical literature but has had little impact upon pedagogic texts.

Question 5 invited teachers to critique a language testing textbook that they had last used. I present each book mentioned in alphabetical order. In discussing this section I do not report numbers of users for each text. This was not the intention of the analysis, and readers who wish to know more about usage statistics should consult J. D. Brown and Bailey (2008). Similarly, the comments on each book are intended to summarize in a pithy way the views of respondents.

Alderson et al. (1995) was said to be clear and informative, particularly for those who are new to the field. "Thoroughly accessible" seemed to sum up the response of users. Although not being too technical, some thought that a disadvantage of the text was that it does not treat any topic in any depth and the statistical analysis needs further explanation.

Bachman (1990) is considered to be a "crucial text" by many respondents, but as a textbook it is also generally thought to be rather too theoretical, with not enough practical examples.

Bachman (2004) and Bachman and Kunnan (2005) were reported by users to provide an excellent hands-on approach to statistics, although conceptually difficult in places. A small number of users reported small mistakes in the text that have not been picked up by proofreaders, and in a text that relies on statistical examples this is seen as critical, as the following comment makes clear: "The misprints were irritating in the workbook as I did not have enough confidence in my own knowledge to know if they were mistakes or if I had gone wrong somewhere."

Bachman and Palmer (1996) was widely considered to be a comprehensive treatment of the field, with good examples taken from actual test development projects. The concept of "test usefulness" was thought to be much more practical in understanding validity issues than treatments in many other texts. However, it was also observed that the text was less concerned with classroom testing and assessment, which is also important for teachers. The projects in Part 3 were also seen as difficult to try out given the time and resource constraints of teachers.

H. D. Brown (2006), although widely used, was considered to be a very general text, in need of supplementing with other reading. Nevertheless, users believed the topic coverage to be good.

J. D. Brown (2005) is generally seen to be accessible, well written, and clear. A major strength is believed to be the straightforward guidelines offered to practitioners, and a number of users praised the statistical examples for Excel. Many commented on the appropriate balance between theory and practice.

Davidson and Lynch (2002) is a popular book, praised by users for the practical approach to developing and using test specifications as a basis for test development. A number of commentators would have liked the book to contain more examples, drawn from a wider range of test development projects.

Fulcher and Davidson's (2007) "triplet design" was said to be user friendly. It was reported that one strength of the book is that it "doesn't insult the reader's intelligence"—but a downside of the level of the text is that it takes considerable time to digest. Some users thought the text would be more relevant to teachers who already had some background in testing and assessment and may be too challenging for newcomers to the field.

Hughes (1989/2003) is a widely used textbook that was praised for the range of topics and practical examples provided, but some respondents thought that the activities were not very useful in reinforcing the content of the text. Although the book was very accessible (particularly with regard to technical terms), it was surprising how many teachers thought that it was "light on classroom assessment" with too much focus on proficiency testing and standardized testing in general. Some also noticed the absence of more recent topics, such as integrated skills assessment, and the social and ethical aspects of testing.

O'Malley and Pierce (1996) was generally recommended for its treatment of performance testing, although a number of respondents thought that it was too focused on the United States.

The reactions of teachers to these textbooks serve to reinforce the importance of key issues that arise in the analysis of other questions but also add to them in terms of the structure and focus of new materials. In particular, it seems that teachers require

- A text that is not light on theory but explains concepts clearly, especially where statistics are introduced.
- A practical "how-to" guidance, although not prescriptive in nature.
- A balance between classroom and large-scale testing, with illustrations and practical examples drawn from a range of sources and countries.
- Activities that can be reasonably undertaken given the constraints and resources that teachers normally face.

Question 6 was designed to capitalize on the reflections elicited in Question 5, to discover if teachers had ideas for content that were not provided by existing textbooks. Perhaps the most important finding that reinforces the importance of the responses to Question 3 is that teachers consider the design of tests and assessments, especially for use in their own classrooms, as an ongoing design process. What is required, many argued, is a text that explains that process and how it is followed through in both standardized- and classroom-assessment contexts. The second area of interest, for which the new inductive code of "social impact" was used, was direct reference to the social impact of testing, particularly the impact of high-stakes testing upon the lives of learners and the practices of teachers. Slightly more than 10% of respondents wanted to see a treatment of the politics and economics of testing—particularly critiquing the role of test

providers. One typical comment from the data to illustrate this point was, "We need information on testing as an industry, a multi-billion dollar concern and why we have to fight crap when we see it." This was associated with the purpose of testing—why, how, and when testing and assessment should (or should not) take place, and the ethical issues surrounding the use of test scores. A historical context was also seen by some as highly relevant to understanding the emergence of many testing practices, despite its lower rating in Question 4.

In terms of "how-to" guidance, teachers returned to a request for a conceptual explanation of the basic statistics, an introduction to item writing and analysis, the development and use of rating scales for performance tests, and a treatment of how these instruments relate to what we know about language learning and use from applied linguistics and second language acquisition research. This latter area is also a feature of current assessment research by the Second Language Acquisition and Testing in Europe organization (http://www.slate.eu.org/; also see Bartning, Martin, & Vedder, 2010).

Finally, Questions 7 and 8 invited teachers to comment on other features of a textbook that they would like to see developed, along with a final chance to add any other comments that they would like to make. Perhaps not surprisingly, many asked for a set of references that provided useful follow-up reading, a glossary, and practical activities that could be used both by groups and individuals. The most common suggestion was for useful electronic resources with links to interesting websites, activities, and additional information.

AN EXPANDED DEFINITION

This survey has shown that language teachers are very much aware of a variety of assessment needs that are not currently catered for in existing materials designed to improve assessment literacy. The answers to the constructed-response questions in particular are indicative of changes in our understanding of the role of testing in society and a desire to understand more of the "principles" as well as the "how-to" (Davies, 2008), no matter how important the latter may be. Of particular importance is the finding that so many of the respondents recommended that principles be embedded and elucidated within a procedural approach to dealing with the practical nuts-and-bolts matters of building and delivering language tests and assessments—further, that both principles and practice should be discussed within a much wider historical and social context. Finally, it was recommended that this procedural approach should treat large-scale and classroom-based assessment in a much more balanced way.

A working definition of assessment literacy based on these findings may be expressed as

> 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.

This is much wider than definitions in currency, such as "the understanding and appropriate use of assessment practices along with the knowledge of the theoretical and philosophical underpinnings

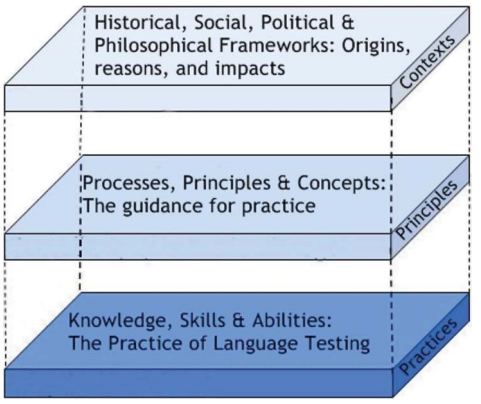

FIGURE 1 Language assessment literacy: An expanded definition (color figure available online).

in the measurement of students' learning" (DeLuca & Klinger, 2010, pp. 419–420), and can be visually represented in Figure 1. However, it is recognized that not all of these components will be essential for all stakeholders.

CONCLUSIONS

In practical terms, the findings and emerging definition suggest an approach to assessment literacy that integrates knowledge, skills, and principles in a procedural text that attempts to balance what will be required for both classroom and normative assessment. The content should be presented within a larger historical, social, political, and ethical framework. Supported with glossary and activities within the text, it may also be linked to an evolving electronic resource that can be used for further reading, research, and study. The outcomes of the research described in this article were used to plan and produce a textbook (Fulcher, 2010) and to further develop a website (http://languagetesting.info) that could be used by students to explore language testing themes, and to see how language testing policy and practice impact upon the world, and individuals. New additions designed explicitly to link testing practice to larger issues in society were a language testing search engine that automatically detects language testing stories from the world press and updates the news on the fly, and a set of language testing scenarios with ideas for group study and project work.

There are a number of limitations to this research that need to be acknowledged, and may be taken into account in future empirical investigations into assessment literacy. The first two of these are related to the nature of the sample used in the study. The second two concern instrumentation, and the last is epistemological.

First, and perhaps most important, we have made reference to the fact that the sample was self-selecting and that the respondents are therefore highly likely to believe that language testing and assessment is an important subject for teachers. The advantage is that these respondents have

useful, thoughtful, and relevant comments to make. However, the disadvantage is that they are likely to think that all topics within language testing are important. This may account for the uniformly high responses to the closed-response items on the survey, and the lack of variation both between and within subgroups of the sample. It therefore seems unlikely that quantitative approaches will yield useful information unless it is possible to conduct research with genuinely random samples, so that they more accurately reflect the range of experience, views, and backgrounds of a known population of teachers. Although this problem does not detract from the richness of the data gathered, and its usefulness in constructing an empirical basis for the description of assessment literacy, it must be acknowledged that further aspects are likely to emerge from the study of a more representative sample.

Another limitation associated with the nature of the sample is the fact that it has not been possible to differentiate between subgroups of the sample, particularly in relation to teachers whose knowledge and understanding of language testing is more cursory, through gradations toward those with extensive knowledge and experience. It is likely that teachers at different stages of assessment literacy acquisition will have different needs, but it has not been possible to identify such differences in this research. The solution, once again, involves using a much more representative sample acquired through randomization.

A potential problem with Question 2 that was not identified during piloting relates to the fact that it does not allow respondents to say whether they have studied language testing and assessment as a separate course or embedded within a more general pedagogy class. These are arguably two very different contexts that may impact upon results. Future studies may wish to consider whether these should be distinguished.

The fourth limitation relates to the construction of the a priori categories for analysis, which grew out of the quantitative analysis. The value and usefulness of this is directly related to the questions that were included on the survey, and for practical purposes these must be limited to numbers that respondents can be expected to answer within a limited time frame. Some of the questions therefore "bundled together" practices and concepts that may usefully have been distinguished. For example, specific needs or practices related to classroom assessment, such as peer and self-assessment, were not separately identified. It is not therefore possible to estimate just how important these practices are to the teachers in the sample. However, for practical purposes it was assumed that these were elements of "classroom practice" and included in the new training materials.

The final limitation relates to the analysis of the qualitative data. Although care was taken to code responses carefully, and to generate additional inductive codes that provided a reasonable interpretation of salient comments, such analysis can always be challenged as "subjective." One way to defend against such challenges is to engage multiple coders and calculate intercoder reliability indexes. This was not possible in this research, due to limitations of time and resources. It is therefore necessary for the reader to consider the interpretations provided and the illustrative quotations in an evaluation of whether inferences and explanations appear reasonable and practically useful for the intended purpose.

Despite these limitations, it is argued that this research, the findings of the survey, the expanded definition, and the resulting materials contribute toward the field's evolving concept of "assessment literacy." The research may also show how a research base can be constructed and used to support pedagogic decisions in the structuring and delivery of materials for teaching language testing and improving assessment literacy.

ACKNOWLEDGMENTS

The research reported in this article was funded by a Leverhulme Research Fellowship (http://www.leverhulme.ac.uk/). I am also grateful to the anonymous reviewers for their generosity in providing stimulating constructive feedback that has greatly improved this article.

REFERENCES

Alderson, J. C., Clapham, C., & Wall, D. (1995). *Language test construction and evaluation.* Cambridge, England: Cambridge University Press.

American Federation of Teachers. (1990). *Standards for teacher competence in educational assessment of students*. Washington DC: Author. Available from http://www.unl.edu/buros/bimm/html/article3.html

Assessment Reform Group. (2002). *Assessment for Learning: 10 Principles. Research Based Principles to Guide Classroom Practice*. London, UK: Author. Available from http://languagetesting.info/features/afl/4031afl principles.pdf

Bachman, L. (1990). *Fundamental considerations in language testing*. Oxford, UK: Oxford University Press.

Bachman, L. (2004). *Statistical analyses for language assessment*. Cambridge, UK: Cambridge University Press.

Bachman, L., & Kunnan, A. (2005). *Statistical analyses for language assessment: Workbook and CD ROM*. Cambridge, UK: Cambridge University Press.

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

Bailey, K. M., & Brown, J. D. (1996). Language testing courses: What are they? In A. Cumming & R. Berwick (Eds.), *Validation in language testing* (pp. 236–256). London, UK: Multilingual Matters.

Barting, I., Martin, M., & Vedder, I. (Eds.). (2010). *Communicative proficiency and linguistic development: intersections between SLA and language testing research* (Eurosla Monograph 1). Available from http://eurosla.org/monographs/EM01/EM01home.html

Black, P., & Wiliam, D. (1998). Inside the black box: Raising standards through classroom assessment. *Phi Delta Kappan, 80*.

Brindley, G. (2001). Language assessment and professional development. In C. Elder, A. Brown, E. Grove, K. Hall, N. Iwashita, T. Lumley, . . . K. O'Loughlin (Eds.), *Experimenting with uncertainty. Essays in honour of Alan Davies* (pp. 126–136). Cambridge, UK: Cambridge University Press.

Brindley, G. (2008). Educational reform and language testing. In E. Shohamy & N. Hornberger (Eds.), *Encyclopedia of language and education: Vol. 7, Language testing and assessment* (pp. 365–378). New York, NY: Springer.

Brown, H. D. (2006). *Language assessment: Principles and classroom practice*. New York, NY: Pearson.

Brown, J. D. (2001). *Using surveys in language programs*. Cambridge, UK: Cambridge University Press.

Brown, J. D. (2005). *Testing in language testing programs: A comprehensive guide to English language assessment*. New York, NY: McGraw-Hill.

Brown, J. D. (2010). How are PCA and EFA used in language test and questionnaire development? *Shiken*, *14*(2), 30–35.

Brown, J. D., & Bailey, K. M. (2008). Language testing courses: What are they in 2007? *Language Testing*, *25*, 349–384.

Cheng, L. (2008). Washback, impact and consequences. In E. Shohamy & N. H. Hornberger (Eds.), *Encyclopedia of language and education* (pp. 349–364). New York, NY: Springer.

Child, D. (1990). *The essentials of factor analysis* (2nd ed.). New York, NY: Cassell Educational.

Davidson, F., & Lynch, B. K. (2002) *Testcraft: A teacher's guide to writing and using language test specifications*. New Haven, CT: Yale University Press.

Davies, A. (2008). Textbook trends in teaching language testing. *Language Testing*, *25*, 327–348.

DeLuca, C., & Klinger, D. A. (2010). Assessment literacy development: identifying gaps in teacher candidates' learning. *Assessment in Education: Principles, Policy & Practice 17*, 419–438.

Extra, G., Spotti, M., & Van Avermaet, P. (Eds.). (2009). *Language testing, migration and citizenship: Cross-national perspectives on integration regimes*. London, UK: Continuum.

Fertig, M. (2003). *Who's to blame? The determinants of German students' achievement in the PISA 2000 study* (Social Science Research Network: ISA Discussion Paper 739).

Fulcher, G. (2009). Test use and political philosophy. *Annual Review of Applied Linguistics*, *29*, 3–20.

Fulcher, G. (2010). *Practical language testing*. London, UK: Hodder Education.

Fulcher, G. (in press). Language testing and philosophy. In A. Kunnan (Ed.), *Companion to language assessment*. London, UK: Wiley-Blackwell.

Fulcher, G., & Davidson, F. (2007). *Language testing and assessment: An advanced resource book*. London, UK: Routledge.

Hasselgreen, A., Carlsen, C., & Helness, H. (2004). *European Survey of Language Testing and Assessment Needs. Part 1: General findings*. Gothenburg, Sweden: European Association for Language Testing and Assessment. Available from http://www.ealta.eu.org/documents/resources/survey-report-pt1.pdf

Hogan-Brun, G., Mar-Lolinero, C., & Stevenson, P. (Eds.). (2009). *Discourses on language and integration: Critical perspectives on language testing regimes in Europe*. Amsterdam, the Netherlands: John Benjamins.

Hughes, A. (2003). *Language testing for teachers*. Cambridge, UK: Cambridge University Press. (Original work published 1989)

Huhta, A., Hirvalä, T., & Banerjee, J. (2005). *European Survey of Language Testing and Assessment Needs. Part 2: Regional findings*. Gothenburg, Sweden: European Association for Language Testing and Assessment. Available from http://users.jyu.fi/~huhta/ENLTA2/First_page.html

Inbar-Lourie, O. (2008a). Constructing a language assessment knowledge base: A focus on language assessment courses. *Language Testing*, *25*, 385–402.

Inbar-Lourie, O. (2008b). Language assessment culture. In E. Shohamy & N. Hornberger (Eds.), *Encyclopedia of language and education: Vol. 7, Language testing and assessment* (pp. 285–300). New York, NY: Springer.

Jin, Y. (2010). The place of language testing and assessment in the professional preparation of foreign language teachers in China. *Language Testing*, *27*, 555–584.

Kunnan, A. J. (2009). Politics and legislation in citizenship testing in the United States. *Annual Review of Applied Linguistics*, *29*, 37–48.

Lantolf, J. P. (2009). Dynamic assessment: The dialectic integration of instruction and assessment. *Language Teaching*, *42*, 355–368.

Lantolf, J. P., & Poehner, M. E. (2008). Dynamic assessment. In E. Shohamy & N. H. Hornberger (Eds.), *Encyclopedia of language and education, Vol.7, Language testing and assessment* (pp. 273–284). New York, NY: Springer.

Long, M. (1991). Focus on form: A design feature in language teaching methodology. In K. De Bot, R. Ginsberg, & C. Kramsch (Eds.), *Foreign language research in cross-cultural perspective* (pp. 39–52). Amsterdam, the Netherlands: John Benjamins.

Malone, M. (2008). Training in language assessment. In E. Shohamy & N. Hornberger (Eds.), *Encyclopedia of language and education: Vol. 7, Language testing and assessment* (pp. 225–239). New York, NY: Springer.

McGaw, B. (2008). The role of the OECD in international comparative studies of achievement. *Assessment in Education: Principles, Policy and Practice*, *15*, 223–243.

McNamara, T. (2008). The socio-political and power dimensions of tests. In E. Shohamy & N. Hornberger (Eds.) *Encyclopedia of language and education; Volume 7: Language testing and assessment* (pp. 415–427). New York, NY: Springer.

McNamara, T., & Roever, C. (2006). *Language testing: The social dimension*. London, UK: Blackwell.

McNemar, Q. (1951). The factors in factoring behavior. *Psychometrika*, *16*, 353–359.

Miles, M. B., & Huberman, A. M. (1994). *Qualitative data analysis* (2nd ed.). London, UK: Sage.

Nietzsche, F. (2005). *The Anti-Christ, ecce homo, twilight of the idols* (A. Ridley & J. Norman, Eds.). Cambridge, UK: Cambridge University Press.

O'Malley, J., & Pierce, L. V. (1996). *Authentic assessment for English language learners: Practical approaches for teachers*. Boston, MA: Addison-Wesley.

Plake, B., & Impara, J. C. (1993). Assessment competencies of teachers: A national survey. *Educational Measurement: Issues and Practice*, *12*(4), 10–12.

Plake, B., & Impara, J. C. (1996). Teacher assessment literacy: What do teachers know about assessment? In G. D. Phye (Ed.), *Handbook of classroom assessment: Learning, achievement and adjustment* (pp. 53–68). New York, NY: Academic Press.

Rea-Dickins, P. (2006). Currents and eddies in the discourse of assessment: a learning-focused interpretation. *International Journal of Applied Linguistics*, *16*, 163–188.

Rea-Dickins, P. (2008). Classroom-based language assessment. In E. Shohamy & N. H. Hornberger (Eds.) *Encyclopedia of language and education, Vol.7, Language testing and assessment* (pp. 257–271). New York, NY: Springer.

Shepard, L. (2000). The role of assessment in a learning culture. *Educational Researcher*, *29*(7), 4–14.
Shohamy, E. (2001). *The power of tests*. Harlow, UK: Longman/Pearson.
Spolsky, B. (1995). *Measured words*. Oxford, UK: Oxford University Press.
Slade, C., & Möllering, M. (Eds.). (2010). *From migrant to citizen: Testing language, testing culture*. Basingstoke, UK: Palgrave Macmillan.
Stiggins, R. (1991). Assessment literacy. *The Phi Delta Kappan*, *72*, 534–539.
Stiggins, R. (1997). *Student-centered classroom assessment*. Upper Saddle River, NJ: Prentice Hall.
Taylor, L. (2009). Developing assessment literacy. *Annual Review of Applied Linguistics*, *29*, 21–36.
Wall, D. (2005). *The impact of high-stakes examinations on classroom teaching: A case study using insights from testing and innovation theory*. Cambridge, UK: Cambridge University press.
Wall, D. (2012). Washback. In G. Fulcher & F. Davidson (Eds.), *The Routledge handbook of language testing* (pp. 79–92). London, UK: Routledge.
Walters, F. S. (2010). Cultivating assessment literacy: Standards evaluation through language-test specification reverse engineering. *Language Assessment Quarterly*, *7*, 317–342.

APPENDIX

The Survey

Q1 Are you a

[Designed to confirm that respondents are, or have been, language teachers]

Q2 When you last studied language testing, which parts of your course you thought were most relevant to your needs?

Q3 Are there any skills that you still need?

Q4 Please look at each of the following topics in language testing.

For each one please decide whether you think this is a topic that should be included in a course on language testing.

Indicate your response as follows:

5 = essential
4 = important
3 = fairly important
2 = not very important
1 = unimportant

A. History of Language Testing ○ 1 ○ 2 ○ 3 ○ 4 ○ 5
B. Procedures in language test design ○ 1 ○ 2 ○ 3 ○ 4 ○ 5
C. Deciding what to test ○ 1 ○ 2 ○ 3 ○ 4 ○ 5
D. Writing test specifications/blueprints ○ 1 ○ 2 ○ 3 ○ 4 ○ 5
E. Writing test tasks and items ○ 1 ○ 2 ○ 3 ○ 4 ○ 5
F. Evaluating language tests ○ 1 ○ 2 ○ 3 ○ 4 ○ 5
G. Interpreting scores ○ 1 ○ 2 ○ 3 ○ 4 ○ 5
H. Test analysis ○ 1 ○ 2 ○ 3 ○ 4 ○ 5
I. Selecting tests for your own use ○ 1 ○ 2 ○ 3 ○ 4 ○ 5

J. Reliability ○ 1 ○ 2 ○ 3 ○ 4 ○ 5
K. Validation ○ 1 ○ 2 ○ 3 ○ 4 ○ 5
L. Use of statistics ○ 1 ○ 2 ○ 3 ○ 4 ○ 5
M. Rating performance tests (speaking/writing) ○ 1 ○ 2 ○ 3 ○ 4 ○ 5
N. Scoring closed-response items ○ 1 ○ 2 ○ 3 ○ 4 ○ 5
O. Classroom assessment ○ 1 ○ 2 ○ 3 ○ 4 ○ 5
P. Large-scale testing ○ 1 ○ 2 ○ 3 ○ 4 ○ 5
Q. Standard setting ○ 1 ○ 2 ○ 3 ○ 4 ○ 5
R. Preparing learners to take tests ○ 1 ○ 2 ○ 3 ○ 4 ○ 5
S. Washback on the classroom ○ 1 ○ 2 ○ 3 ○ 4 ○ 5
T. Test administration ○ 1 ○ 2 ○ 3 ○ 4 ○ 5
U. Ethical considerations in testing ○ 1 ○ 2 ○ 3 ○ 4 ○ 5
V. The uses of tests in society ○ 1 ○ 2 ○ 3 ○ 4 ○ 5
W. Principles of educational measurement ○ 1 ○ 2 ○ 3 ○ 4 ○ 5

Q5 Which was the last language testing book you studied or used in class?

What did you like about the book? What did you dislike about the book?

Q6 What do you think are essential topics in a book on practical language testing?

Q7 What other features (e.g. glossary/activities etc) would you most like to see in a book on practical language testing?

Q8 Do you have any other comments that will help me to understand your needs in a book on practical language testing?

Q9 How would you rate your knowledge and understanding of language testing?

5 = very good
4 = good
3 = average
2 = poor
1 = very poor

Q10 And now, just a few quick questions about you.

Are you male or female?
○ Female | ○ Male

Q11 What is your age range?

○ Under 20
○ 21 - 25
○ 26 - 30
○ 31 - 35
○ 36 - 40
○ 41 - 45
○ 46 - 50
○ 51 - 55

- 56 - 60
- 61 - 65
- Above 65

Q12 Please select your current educational level

High School Graduate
BA degree
MA degree
Doctorate
Other

Q13 Which is your home country?

Which country do you currently live or study in?

Q14 Which language do you consider your first language?

What other languages do you speak?