Author's accepted version: Benjamin Kremmel (University of Innsbruck) & Luke Harding (Lancaster University). To appear in *Language Assessment Quarterly*.

Towards a comprehensive, empirical model of language assessment literacy across stakeholder groups: Developing the Language Assessment Literacy Survey

Abstract

While scholars have proposed different models of language assessment literacy (LAL), these models have mostly comprised prescribed sets of components based on principles of good practice. As such, these models remain theoretical in nature, and represent the perspectives of language assessment researchers rather than stakeholders themselves. The project from which the current study is drawn was designed to address this issue through an empirical investigation of the LAL needs of different stakeholder groups. Central to this aim was the development of a rigorous and comprehensive survey which would illuminate the dimensionality of LAL and generate profiles of needs across these dimensions. This paper reports on the development of an instrument designed for this purpose: the Language Assessment Literacy Survey. We first describe the expert review and pretesting stages of survey development. Then we report on the results of an exploratory factor analysis based on data from a large-scale administration (N = 1086), where respondents from a range of stakeholder groups across the world judged the LAL needs of their peers. Finally, selected results from the large-scale administration are presented to illustrate the survey's utility, specifically comparing the responses of language teachers, language testing/assessment developers and language testing/assessment researchers.

Introduction

Given the widespread use of language assessments for decision-making across an increasing number of social domains (education, immigration and citizenship, professional certification), it has become vital to raise awareness and knowledge of good practice in language assessment for a wide range of stakeholder groups. Scholars have thus called for the promotion of language assessment literacy (LAL) not only for teachers and assessment developers, the two groups most typically involved with language assessments, but also for score users, policymakers and students (among others) (e.g. Baker, 2016; Deygers & Malone, 2019). For such groups, a heightened awareness of the principles and practice of language assessment would ideally lead to more informed discussion of assessment matters, clarity around good practice in using language assessments, and ultimately more robust decision-making on the basis of assessment data (O'Loughlin, 2013; Pill & Harding, 2013; Taylor, 2009).

Yet it is still unclear what, and how much, different stakeholder groups should know about language assessment in order to perform their specific assessment-related tasks, and to engage in meaningful interpretations and critical discussions about assessment practices (Harding & Kremmel, 2016). Although speculative profiles for different groups have been developed (e.g., Taylor, 2013), there is a gap in our understanding of the perceived LAL needs of the stakeholders themselves, and how these might differ across different roles and professions. At the same time, gauging the needs of different roles and professions requires the development of instruments which can elicit comparable data on these needs across a range of groups; broadening the dimensions of language assessment literacy beyond those typically assumed to be of relevance to teachers or assessment specialists.

The aim of the present paper is to describe the development and initial findings of a large-scale questionnaire – the Language Assessment Literacy Survey – which was designed to address the research gap by gathering data-driven descriptive evidence to support current prescriptive claims for stakeholders' LAL needs. Specifically, we aim to provide empirical backing drawing on survey data to evaluate Taylor's (2013) LAL profiles in terms of the hypothesised dimensions of LAL, and the degree to which LAL may differ across key stakeholder groups. In parallel, the paper provides the first published report on the development and factor structure of the Language Assessment Literacy Survey, an instrument designed for use across different contexts and stakeholder groups.

Background

The notion that separate LAL profiles might exist for different stakeholder groups emerged as LAL research developed and diversified. Early contributions to the assessment literacy literature, both in general education (e.g., Popham, 2006; Stiggins, 1991) and in language assessment (Brindley, 2001; Davies, 2008) concentrated on identifying the components of assessment knowledge and skills primarily required of teachers. This emphasis is still prevalent in more recent research, both in terms of general assessment literacy (e.g. Mertler, 2009; Mertler & Campbell, 2005; Plake, Impara & Fager, 1993) as well as assessment literacy more specific to language teachers (e.g. Lam, 2015; Vogt & Tsagari, 2014). This is not surprising as teachers are at the frontline as designers and users of language assessments and there is thus a clear need for language educators to be "conversant and competent in the principles and practice of language assessment" (Harding & Kremmel, 2016, p. 415). However, the important role of language assessment in decision-making processes across a range of domains, and the diverse nature of stakeholder groups involved in assessment processes, demands a view of LAL that extends beyond a focus on teachers. This was noted by Taylor (2009), who identified that LAL is needed for a wide range of social actors:

> … personnel in both existing and newly established and emerging national examination boards, academics and students engaged in language testing research, language teachers or instructors, advisors and decision makers in language planning and education policy, parents, politicians and the greater public. (p. 25)

If LAL is seen to be required across diverse groups, it follows that individuals in different professional/social roles may have different LAL requirements based on circumstantial requirements; or as Pill and Harding (2013) state: "different levels of expertise or specialization will require different levels of [language assessment] literacy, and different needs will dictate the type of knowledge most useful for stakeholders" (p. 383).

While recent research has provided some backing for the notion of unique LAL needs within specific stakeholder groups (e.g., admissions officers in O'Loughlin, 2013; policy makers in Pill & Harding, 2013; TESOL/applied linguistics lecturers in Jeong, 2013), there is as yet no clear understanding of how differentiated LAL needs might be mapped across such groups. Underpinning this problem is that definitions of LAL – the nature and scope of the construct – have differed widely within the literature (e.g., Brindley, 2001; Davies, 2008; Inbar Lourie, 2008; Fulcher, 2012; Pill & Harding, 2013), and have often not provided sufficient detail to enable a diagnostic approach to identifying unique profiles. In addition, despite some notable exceptions that have yielded useful insights into the LAL needs of teachers (e.g. Fulcher, 2012; Vogt & Tsagari, 2014), many past and current definitions and conceptualizations of LAL remain hypothetical, representing theoretical models devised by language assessment researchers. As a result, our understanding of the extent to which different stakeholder groups have specific LAL needs remains obscure.

An important step towards developing LAL profiles was the shift from more componential views of LAL (e.g., Brindley, 2001; Davies, 2008; Inbar Lourie, 2008), to consideration of developmental trajectories. For example, Fulcher (2012) provides a broad classification of LAL into (a) practical knowledge, (b) theoretical and procedural knowledge, and (c) socio-historical understanding, arguing that practical knowledge provides the foundation of LAL before moving into the more theoretical and principled understandings. Pill and Harding (2013) drew on models from mathematics and science literacy in outlining a continuum of LAL from “illiteracy”, through “nominal literacy”, “functional literacy” and “procedural and conceptual literacy”, to an expert level of knowledge: “multidimensional language assessment literacy” (p. 383).

The notion that LAL may be both multidimensional and developmental paved the way for Taylor (2013), in her summary paper for the special issue of *Language Testing* on language assessment literacy, to merge Pill and Harding’s developmental scale with a synthesized framework of components drawn from recent LAL literature. Taylor suggested that it was important to think about LAL in terms of profiles, which would map-out specific levels of knowledge required across LAL dimensions for different stakeholder groups. Taylor proposed eight dimensions: 1) knowledge of theory, 2) technical skills, 3) principles and concepts, 4) language pedagogy, 5) sociocultural values, 6) local practices, 7) personal beliefs/attitudes, and 8) scores and decision making. Although Taylor was careful not to label this a *model* of LAL, and made clear that the suggestions were speculative, the profiles offered a useful starting point for a more elaborate conceptualization of LAL showing distinct LAL profiles and requirements of different groups. As an illustration, Taylor tentatively drew-up profiles of four key stakeholder groups (Figure 1).

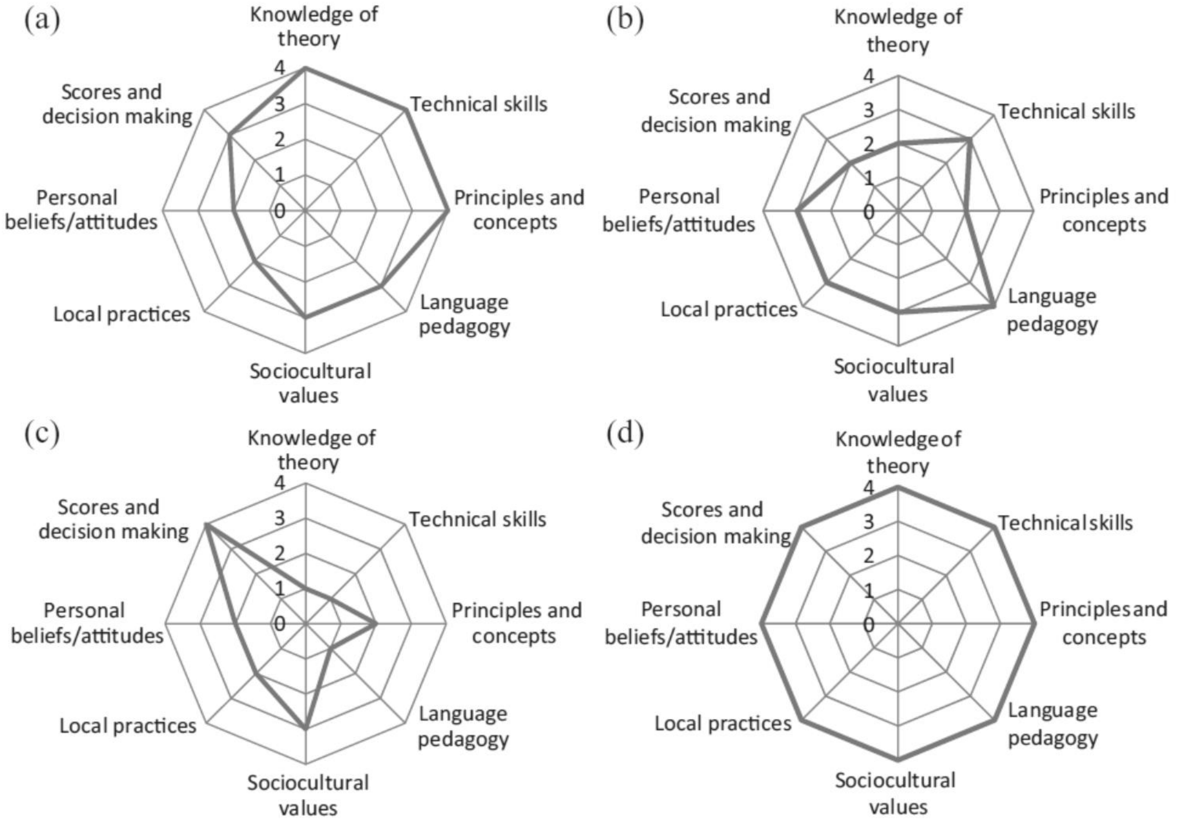

Figure 1: LAL profiles of four stakeholder groups (a=test writers [e.g., test developers], b=classroom teachers, c=university administrators, d=professional language testers [researchers]) (Taylor, 2013, p. 410)

Taylor's notion of LAL profiles has already had significant resonance in the field. In their investigation of the LAL development of 120 Haitian language teachers, Baker and Riches (2017) found the concept useful to track LAL gains, while also making modifications and additions to Taylor's model. Yan, Zhang & Fan (2018) also used the profiles as a point of comparison in a study of language teachers' LAL needs in China. However, the speculative nature of the profiles, the "etic" view they embody, and the need to broaden the profiles to a wider group of stakeholders represents an important gap in LAL research. In addressing these gaps, the present study aimed to elaborate and validate Taylor's profiles by means of a large-scale survey that invited a range of stakeholder groups to assess their needs and identify how important they consider various aspects of LAL for members of their group/profession. Specifically, two research questions were posed:

(1) To what extent are hypothetically different dimensions of language assessment literacy empirically distinct?

(2) To what extent, and in what ways, do the needs of different stakeholder groups vary with respect to identified dimensions?

Method

Instrument development

A number of existing LAL survey instruments have been reported in the research literature – most prominently Fulcher’s (2012) survey, which has been modified for use in numerous research contexts, and the survey used by Vogt & Tsagari (2014) to evaluate assessment literacy across Europe. However, as these surveys were designed primarily for teachers, and therefore for a different purpose to the present instrument, they accordingly may not reflect the full range of assessment-related activities that would be undertaken by a range of different stakeholder groups. Thus, in order to develop a language assessment literacy survey to be used by a range of stakeholders to assess their own groups’ needs, we had two clear guiding aims: (1) the survey would need to be comprehensive, yet feasible to complete among populations where motivation to engage with LAL may be low; and (2) the survey items would need to be intelligible across the wide-range of stakeholder groups suggested by Taylor (2013). This necessitated a multi-stage development process which spanned almost 24 months (see Figure 2).

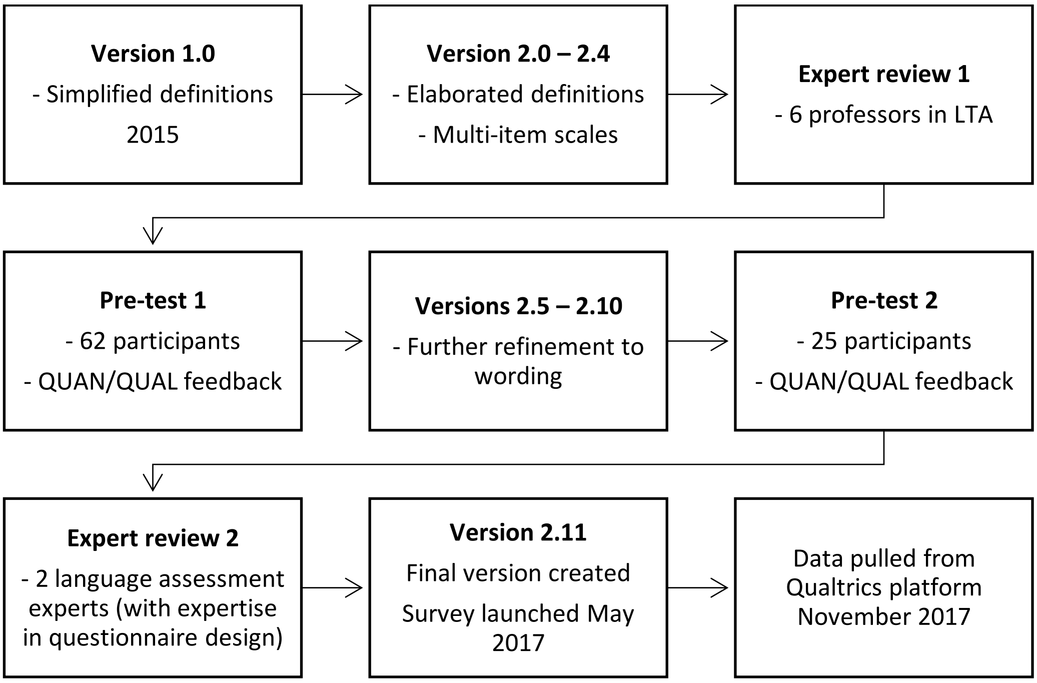

Figure 2: Overview of instrument development process

The starting point for instrument development was Taylor's (2013) hypothesised dimensions of LAL, and specifically the eight components described above. After an initial pilot of a much more basic questionnaire (version 1.0), we began to develop a survey which would consist of multi-item scales for each hypothesised dimension, with the aim of generating a minimum of four items per sub-scale (see Dörnyei & Taguchi, 2010). In order to flesh-out the items in these scales we drew on multiple published sources which presented assessment literacy questionnaires specifically for teachers (Fulcher, 2012; Stiggins, 1991), and brainstormed our own items within the hypothesised categories drawing on the literature surveyed above. The initial survey underwent four revisions between the two researchers (versions 2.0 – 2.3). During this process, in keeping with our original aims, we focused on developing a set of short items, rendered in simple language, with glosses provided where necessary. At the same time, we expanded on the number of hypothesised dimensions (dividing technical skills into three different areas: language assessment construction, language assessment administration/scoring, and language assessment evaluation). We also modified the category labels for the various stakeholders who would be targeted by the survey, moving beyond Taylor's (2013) initial classifications to include professional examiners/raters and test-takers and separating "policy makers" into "policy makers" and "test score users". "Policy makers" we thereby defined as, e.g. 'a [government] official who sets educational goals and assessment policies', and "test score user" as 'e.g., university admissions staff, employers etc. who might use language test scores for decision making'. A further category "parent of a test taker (parent or legal guardian whose child is taking a language test)" was added. All participants saw these exemplifying definitions in the survey so as to clarify distinctions between these groups as much as possible. Finally, we developed an initial five-point response scale for the survey which ranged from 0 to 4, with 0 representing the perceived need for "no knowledge/skill at all" on a particular attribute, and 4 representing the perceived need for "a very high/comprehensive level of knowledge/skill".

The first full version of the survey (v2.3) was then used as the basis for an expert review. This version consisted of 70 items, with a mean of 7 items per dimension (min = 5; max = 12). We recruited six experts, all of whom were professors or senior researchers in the field of language testing and assessment, to complete the survey and comment on (a) anything which appeared odd/out-of-place, (b) any obvious omissions within each domain category, (c) any less relevant items which could be removed, and (d) any other general views on the survey. This process yielded numerous suggestions for changes to the wording of items and to the scale for clarity and cohesion. We adopted these suggestions for version 2.4, which was the first online version of the survey, developed on the Qualtrics platform.

Version 2.4 was used for the first pre-test, which we conducted with 62 participants across a range of stakeholder groups. The pre-test was primarily designed to gather feedback from all targeted stakeholder groups concerning the clarity and comprehensiveness of the survey, thus including a range of voices in the survey design beyond those of the testing experts. During this pre-test we collected quantitative data (respondents' judgements of the clarity of each survey section), as well as qualitative data on the user experience. Comments gleaned from the first pre-test led to several more revisions by the researchers (versions 2.5-2.10) before another pre-test with 25 participants, and a further review by two experts (one with specific experience in questionnaire design) to confirm the suitability of the changes made following the first pre-test. Final changes were implemented in version 2.11, and this version was officially launched in May 2017 (available at: https://tinyurl.com/LALsurvey1).

Instrument format

Survey respondents were first shown a screen which provided basic information about the survey and provided a link to an information sheet about the project. Respondents who chose to continue were then directed to a screen which asked them to select the group(s)/profession(s) that they identified with. Respondents were asked to select all of the identities that applied to them from the following list:

- Language teacher
- Professional examiner and/or rater
- Language assessment/test developer (a professional who creates tests or assessments, writes questions, develops scoring guides, etc.)
- Language assessment/testing researcher (a professional who conducts research on language testing/assessment matters)
- Policy-maker (a [government] official who sets educational goals and assessment policies)
- Test score user (e.g. university admissions staff, employers, etc. who might use language test scores for decision making)
- Test taker (language learner who might need to take a language test)
- Parent of a test taker (parent or legal guardian whose child is taking a language test)

The rationale for asking respondents to choose a range of identities was that pre-testing had shown many potential respondents held multiple identities (e.g., they were both language teachers and professional examiners, or language assessment researchers and test developers). Allowing respondents to indicate all of their roles/professional identities was therefore seen as useful both in terms of data collection and making the experience less frustrating for users. The next screen asked respondents to select *one* of those identities to focus on for the purposes of the survey. Those who selected “language teacher” on the first screen were diverted to another screen to indicate the level at which they taught (primary, secondary, further/higher, adult), and all professionals (e.g., teachers, test developers, etc.) were asked to indicate what sort of institution they worked at (government, private, non-profit, educational institution, other). A final question, after respondents had selected their focal role/profession, asked how experienced respondents perceived themselves to be in their current role: novice, competent or expert (based on a simplified version of the Dreyfus model of expertise, Dreyfus & Dreyfus, 1980).

Having completed the preliminary questions, respondents were taken to the main set of items designed to identify how important they considered various aspects of LAL for members of their role/profession. The survey contained 71 items, which had been written to relate to ten hypothesized dimensions (see Table 1). The full set of 71 items in the administered version of the survey is provided in Appendix 1:

Table 1: Hypothesized items and related item numbers

Hypothesized dimensions	Item numbers (see Appendix 1)
Knowledge of theory	26, 27, 28, 33, 41, 42
Principles and concepts	31, 32, 40, 43, 44
Language pedagogy	1, 2, 3, 4, 5, 6, 7, 8, 17, 19, 21, 23, 24, 25
Impact and social values	18, 22, 29, 34, 35, 36, 37
Local practices	11, 12, 13, 14, 38, 39
Personal beliefs/attitudes	45, 46, 47, 48
Scores and decision-making	9, 10, 15, 16, 20, 30,
Technical skills (A) – Constructing language assessments	54, 57, 58, 59, 60, 62, 63, 68, 69, 70, 71
Technical skills (B) – Administering/scoring language assessments	53, 55, 56, 61, 67
Technical skills (C) – Evaluating language assessments	49, 50, 51, 52, 64, 65, 66

Given the complexity and length of the survey, and following feedback from expert review and pre-testing, items were organised in the survey so that those items formulated in syntactically similar ways were grouped together. The rationale for this was so respondents would not become confused by multiple switches in syntactic structures across different question stems, and would be able to complete the survey more efficiently. Responses on each item were made on a five-point scale (respondents clicked in a button on the online survey):

How knowledgeable do people in your chosen group/profession **need to** be about each aspect of language assessment below? Please respond according to the following scale:
0 = not knowledgeable at all
1 = slightly knowledgeable
2 = moderately knowledgeable
3 = very knowledgeable
4 = extremely knowledgeable

This scale had been developed and modified during pre-testing, and provided the most useful way of assessing the perceptions of needs among different roles/professional groups. An almost identical question was used for a set of items (grouped together) which referred to skills rather than types of knowledge (see Appendix 1).

Respondents who completed the 71 items were asked to provide a confidence rating for their responses using a sliding scale (0% to 100% confident), and to complete biodata questions eliciting: gender, age, years of experience in role/profession, country of residence, main language used in professional/learning role. A space for open-ended comments was also provided. Respondents were finally asked if they would like to continue on to provide a self-assessment of their own knowledge/skill on the same set of items (this analysis is not within the scope of the current paper).

Main trial sample

We did not use a probability sampling approach in the main trial because the size of our target population for each category was unknown (e.g., there is no reliable data on the number of language teachers worldwide). We also faced a challenge in gaining access to members of the various stakeholder groups and encouraging them to complete the survey. This was partly because networks of language professionals (e.g., teachers, examiners) working within organizations are geographically dispersed and difficult to reach. For that reason, we implemented a mixture of non-probability sampling techniques—purposive (maximum variation) sampling and snowball sampling—with the aim of recruiting a large number of participants to provide a diverse and comprehensive range of respondents within the various participant categories.

Sampling was conducted using a variety of methods. We first placed an invitation to participate (with a weblink) on various professional discussion lists, such as LTEST-L and the EALTA mailing lists, and also encouraged those who saw the email to forward the invitation to others within their networks who fit one of the stakeholder categories. We then posted several messages on Twitter during the recruitment period (May-November 2017) and on a range of IATEFL Facebook groups related to language teaching. We encouraged members of these groups to share the survey link with other local professional networks via social media. This was a particularly useful way of accessing language teachers in various places around the world where we did not have direct contacts. We also contacted specific individuals who would have influence within professional networks in particular countries—in order to increase the heterogeneity of the sample within each role/professional group—sending email invitations which could be circulated across discussion group lists to which we did not have access ourselves. We recognise that non-probability sampling techniques create the potential for biases in the final results, with concentrations among certain network groupings, and with the type of self-selection bias that is inherent in any non-probability sampling. At the same time, our decision to administer an online survey, and to use almost entirely web-based recruitment techniques across social media and discussion fora, meant that we were able to tap into a large and geographically diverse sample. We also note that, based on this method, our sampling of language assessment researchers and language test developers in particular represents the largest sampling of those two groups in the literature to date.

Data for the current paper were exported from Qualtrics on 16 November 2017. By that time, the survey had been live for six months. The exported data showed that 2,419 surveys had been started. However, because of the nature of online survey response, we took a conservative approach in cleaning the data, removing (a) any survey which was incomplete (e.g., where the respondent had stopped prior to filling-in the biodata at the end of the survey), and (b) any survey where the respondent had indicated < 50% confidence in their own responses. Through this cleaning process we removed 1,333 surveys[i], resulting in a final sample of *n* = 1,086.

The final set of respondents were spread over 77 different countries; the distribution of the survey and related frequency of response is visualized in Figure 3. Three countries provided around 50% of the responses in the dataset: China (*n*=231), the United Kingdom (*n*=125) and the United States (*n*=116). This was not surprising given the high concentrations of language assessment activities (whether test development and examining, or test use) in these three contexts.

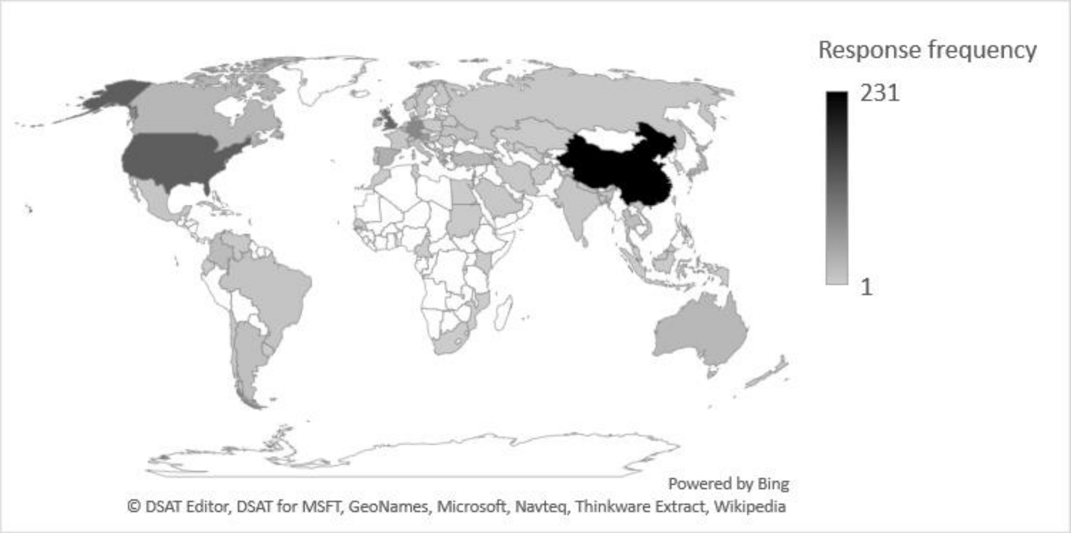
Response frequency
231
1
Powered by Bing
© DSAT Editor, DSAT for MSFT, GeoNames, Microsoft, Navteq, Thinkware Extract, Wikipedia

Figure 3: Distribution of survey and frequency by country

The survey respondents also comprised a range of roles (see Table 2), though with a heavy skew towards language teachers (645), followed by language assessment/test-developers (198) and language assessment/testing researchers (138).

Table 2: Respondent professions/roles

	f	**%**
Language teacher	645	59.4%
Language assessment/test developer	198	18.2%
Language assessment/testing researcher	138	12.7%
Professional examiner/rater	42	3.9%
Test-taker	30	2.8%
Test score user	13	1.2%
Policy-maker	13	1.2%
Parent/legal guardian of a test-taker	7	0.6%

Other sample characteristics relating to gender and age of respondents are provided in Table 3 below.

Table 3: Respondent characteristics

		f	%
Gender	Female	750	69.06%
	Male	309	28.45%
	Other	2	.18%
	Prefer not to say	20	1.84%
	No response	5	.46%
Age	< 20	4	.37%
	21-30	131	12.06%
	31-40	336	30.94%
	41-50	282	25.97%
	51-60	216	19.89%
	> 61	97	8.93%
	Not stated	20	1.84%

Results and discussion

RQ1: To what extent are hypothetically different dimensions of language assessment literacy empirically distinct?

We conducted an exploratory factor analysis (EFA) in order to determine whether (a) there was any empirical basis for the separability of factors in the model of language assessment literacy, and (b) if so, to determine what those factors were. EFA was chosen over a confirmatory factor analysis (CFA) as our intention was firstly to explore the dimensionality of LAL; we had a general idea of what we might find based on Taylor's theoretical profiles, however there was no existing empirical research on which to base our hypotheses. More practically, we also wanted to explore ways to reduce the size of the survey for future research and use. We are aware of the limitations of factor analytic approaches (e.g. van der Eijk & Rose, 2015), however the heavy bias in the sample towards one stakeholder group (teachers) and the fact that the sample sizes of other groups were too small to meaningfully employ psychometrically more sophisticated Mokken analyses (Mokken, 1971), rendered an EFA approach the most feasible.

The EFA analysis was conducted in SPSS (version 23). Before beginning the analysis, the dataset was inspected to ensure that it met the assumptions for factor analysis. The sample size was large enough both with respect to recommendations in the literature concerning absolute size, and item-to-participant ratio, which was approximately 1:15 (see Loewen & Gonulal, 2015). Inspection of an initial correlation matrix of all items showed a high majority of correlations > .30. The Kaiser-Meyer-Olkin (KMO) Measure of Sampling Adequacy and Bartlett's test of sphericity were also calculated. The results (KMO = .976; Bartlett's test of sphericity: p < .001) indicated that factor analysis would be an appropriate method.

We conducted the EFA using principal axis factoring with a direct oblimin rotation. The reason for using principal axis factoring (as opposed to a principal components analysis) was because we aimed to detect the latent constructs underlying response patterns, and not simply to apply factor analysis for data reduction. We chose direct oblimin rotation – an oblique rotation – because we expected that there would be relatively high correlations between the factors which emerged, given that expertise may develop in related ways across the dimensions.

The first run of the factor analysis suggested either a nine- or ten-factor solution. There were ten factors with eigenvalues exceeding 1. However, inspection of the scree plot showed a very small change between the values of 9 and 10, suggesting that 9 factors should be retained. The analysis was re-run with different extraction values specified (including lower values), however the 9-factor solution provided the clearest and most meaningful initial pattern matrix. We then explored those items which had low factor loadings, or which cross-loaded on two or more factors. Here, given the high number of items we began with, we took a relatively conservative approach, removing items with loadings of < .35, or items which loaded on more than one factor at > .35. As we removed items through an iterative process we re-ran the analysis to check that the 9-factor solution held. The outcome of this process was the removal of 21 items, leaving a final collection of 50 items. Every removed item was considered closely, and in all cases there was a clear justification for removal: most often, removed items were, on reflection, ambiguously worded or conceptually similar to another item. The list of removed items is shown in Appendix 2.

The final eigenvalues are shown in Table 4 below, indicating that the final 9-factor solution explained 73.1% of the variance in responses.

Table 4: Eigenvalues for 9-factor solution

	Initial Eigenvalues			**Extraction Sums of Squared Loadings**		
Factor	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	22.065	44.129	44.129	21.755	43.509	43.509
2	4.634	9.267	53.397	4.346	8.691	52.201
3	2.242	4.485	57.882	1.880	3.759	55.960
4	1.840	3.680	61.561	1.549	3.098	59.059
5	1.317	2.634	64.196	1.060	2.121	61.179
6	1.259	2.518	66.713	.979	1.959	63.138
7	1.134	2.269	68.982	.866	1.731	64.869
8	1.040	2.079	71.061	.760	1.519	66.388
9	1.013	2.026	73.088	.671	1.343	67.731

The full rotated pattern matrix is shown in Appendix 3. Following the analysis, the items making up the identified factors were scrutinised to identify their commonalities, and to develop labels for each of the nine dimensions. Factor 1 contained 14 items each related to the processes of constructing language assessments, training others within assessment development contexts, and administering assessments. This factor was labeled *Developing and administering language assessments*. Each of the six items in Factor 2 clearly related to the use of assessments in teaching and learning contexts; this factor was labeled *Assessment in language pedagogy*. Factor 3 also contained six items, each of which related to local practices, but also to policy (e.g., item 22 “how assessments can be used to enforce social policies”). This factor was labeled *Assessment policy and local practices*. Factor 4 contained the four items initially hypothesized to relate to *Personal beliefs and attitudes*; this label was retained. Factor 5 included five items relating to the use of statistics or other methods for analyzing language assessments, labeled *Statistical and research methods*. Factor 6 included four items which related to assessment principles (validity and reliability) as well as score interpretation, labeled *Assessment principles and interpretation*. Factor 7 contained five items which all related to aspects of language, labeled *Language structure, use and development*. Factor 8 contained four items, all initially hypothesized to be related to language pedagogy, but which specifically referred to the preparation of learners for assessments, and the effects of assessment on teaching and learning. This factor was labeled *Washback and preparation*. Finally, Factor 9 included three items which concerned grading processes. This was labeled *Scoring and rating*.

Cronbach’s alpha was calculated to investigate the reliability of each subscale. The final labels, the corresponding items their reliability indices are shown in Table 5.

Table 5: The nine factors of LAL as represented in the final version of the LAL survey

		Item numbers	α
Factor 1	Developing and administering language assessments	62, 68, 61, 63, 64, 66, 70, 69, 65, 60, 67, 58, 59, 17	.96
Factor 2	Assessment in language pedagogy	8, 7, 6, 5, 1, 21	.89
Factor 3	Assessment policy and local practices	12, 11, 38, 14, 39, 22	.88
Factor 4	Personal beliefs and attitudes	46, 47, 45, 48	.93
Factor 5	Statistical and research methods	50, 49, 51, 52	.95
Factor 6	Assessment principles and interpretation	32, 31, 3, 10	.85
Factor 7	Language structure, use and development	28, 27, 26, 29, 33	.85
Factor 8	Washback and preparation	24, 25, 23, 19	.87
Factor 9	Scoring and rating	56, 55, 53	.85

In answer to RQ1, the findings of factor loadings described above suggested that there are nine empirically distinct, separable dimensions of LAL. Further, these dimensions represent an extension and a modification of both Taylor's (2013) initial framework, and our own hypothesized dimensions based on Taylor's work. The evolution of these dimensions across the three stages – initial framework, hypothesized dimensions, data-driven factor structure – is summarized in Table 6.

Table 6: Comparison of dimensions from literature with data-driven factor structure

Taylor's (2013) domains	Hypothesized dimensions	Data-driven factor structure
Knowledge of theory	Theoretical knowledge about language and language learning	Factor 7: Language structure, use and development
Technical skills	(A) Language assessment construction	Factor 1: Developing and administering language assessments
	(B) Language assessment administration/scoring	
	(C) Language assessment evaluation	Factor 5: Statistical and research methods
Principles and concepts	Principles and concepts	Factor 6: Assessment principles and interpretation

Language pedagogy	Language pedagogy	Factor 2: Assessment in language pedagogy
		Factor 8: Washback and preparation
Sociocultural values	Impact and sociocultural values	Factor 3: Assessment policy and local practices
Local practices	Local practices	
Personal beliefs/attitudes	Personal beliefs/attitudes	Factor 4: Personal beliefs and attitudes
Scores and decision making	Scores and decision making	Factor 9: Scoring and rating

Several points are notable. First, Taylor’s dimension “Knowledge of theory”, emerges as a more clearly defined category, with a focus on language and linguistic knowledge. Second, knowledge related to assessment construction and administration appear to be highly-related. This is perhaps not surprising as the administration-related items here refer more to higher-level planning (e.g., developing policy around accommodations) than aspects such as invigilation. Third, statistical and research methods appear to comprise a distinct dimension from other technical design-related skills (as we hypothesized at the item development stage). Fourth, Taylor’s (2013) dimensions “Sociocultural values” and “local practices” appear to form one factor and may be better conceptualized as one combined dimension. This has intuitive appeal, as sociocultural values, which are usually context-dependent, will generally have some impact on practices in local contexts. Finally, “washback and preparation” functions as a standalone dimension, suggesting that concerns around washback may have broad applicability across stakeholder groups (see next section).

RQ2: To what extent, and in what ways, do the needs of different stakeholder groups vary with respect to identified dimensions?

To address RQ2, we generated mean scores on each dimension for key stakeholder group to create LAL profiles. While a detailed analysis and discussion of all stakeholder groups is beyond the scope of this paper (see Harding & Kremmel, in preparation), we sought to illustrate the utility of the survey data by profiling and comparing the three largest stakeholder groups in our sample: language test/assessment (LTA) developers, language testing/assessment (LTA) researchers and language teachers. A comparative LAL needs profile is shown in Figure 4.

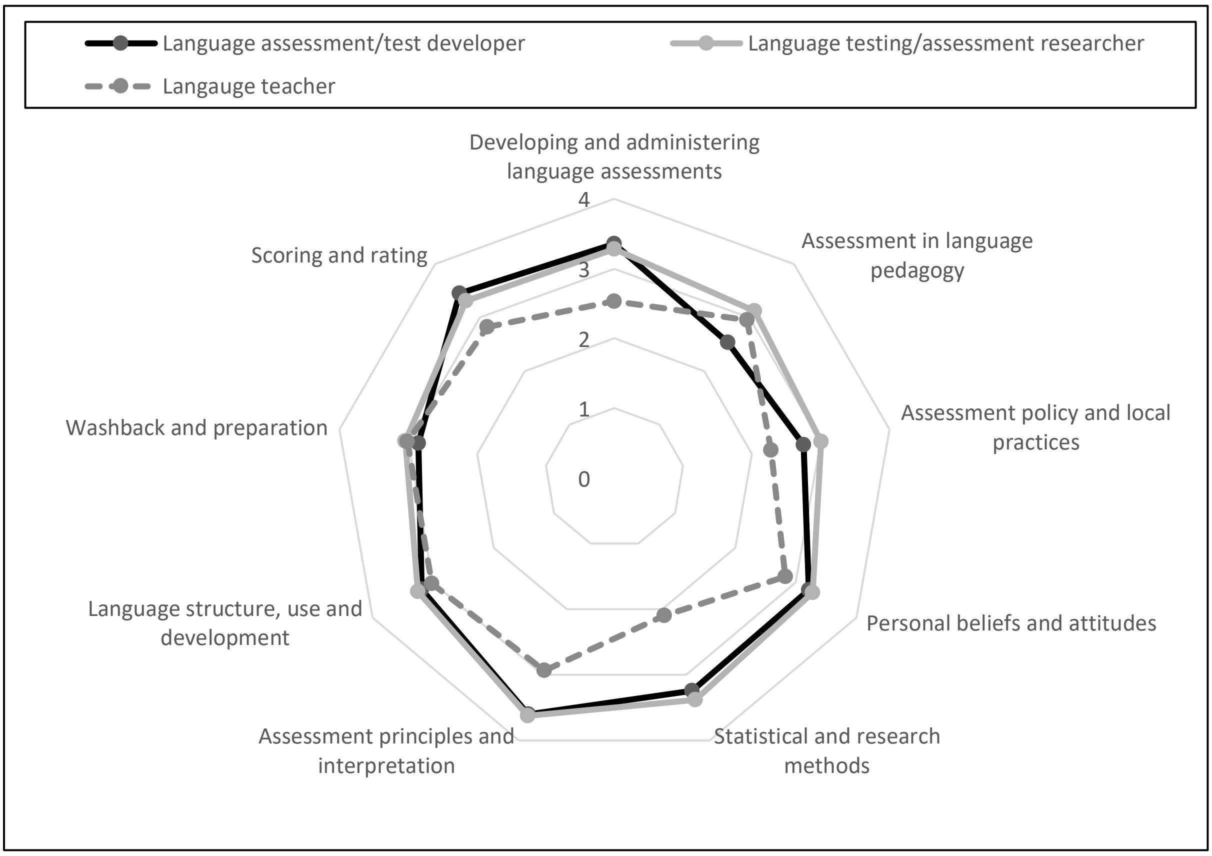

Figure 4: LAL needs profile of three key stakeholder groups: language test/assessment developers (*n*=198); language testing/assessment researchers (*n*=138); language teachers (*n*=645) (note that this is a summary of the perceived needs of respective stakeholder groups rather than their actual competence in these dimensions)

Figure 4 illustrates that language teachers in the sample perceived their role as requiring a reasonably balanced LAL profile, with means on most dimensions around Level 3: "very knowledgeable/skilled" (with the exception of *Developing and administering language assessments*, *Statistical and research methods* and *Assessment policy and local practices* – see discussion below). A table form summary of Figure 4 can be found in Appendix 4. The profile of LTA developers appears to be similarly well-rounded, but noticeably more expansive than the language teacher profile (with the exception of *Assessment in language pedagogy*). Here, most dimensions sit between Level 3 and Level 4, that is, between very and extremely knowledgeable/skilled. The profile of LTA researchers mimics that of language test developers, although with more balance across the nine dimensions, reflecting the notion that LTA researchers (typically university academics) hold complex roles, including evaluating assessment, teaching about assessment, and potentially developing language assessments.

A mixed between-within subjects (3 x 9) ANOVA was initially run to test whether differences in ratings across the three groups were meaningful, with "professional group" included as the between-subjects factor, and "LAL dimension" as the within-subjects factor. However, Levene's test of equality of error variances indicated that the data violated the assumption of homogeneity of variance (potentially problematic given the unequal sample sizes of the three professional groups), and so a non-parametric alternative was deemed more appropriate. Therefore, a series of Kruskal-Wallis tests were run for each of the nine LAL dimensions separately, with professional group as the independent variable. Due to multiple comparisons being made, a Bonferroni correction was applied to the p-value with a new threshold of .006 set. A significant difference ($p < .001$) was observed in mean ranks between groups across all nine dimensions except for *Washback and preparation*. Pairwise between-group comparisons on the remaining eight dimensions were then explored with a series of post-hoc Mann-Whitney U tests, with Bonferroni adjustments applied again ($p = 0.05/24 = .002$). The results (z score for each comparison, effect size [r] and p-value) of the post-hoc tests are shown in Table 7.

Table 7: Pairwise comparisons by three selected professional groups on eight LAL dimensions (asterisk denotes finding significant at *p* = .002).

LAL dimension	Pairwise comparison	*Z*	*r*	*p*
Developing and administering language assessments	LTA developers & LTA researchers	-1.19	-0.06	.235
	LTA developers & language teachers	-12.26	-0.42	.000*
	LTA researchers & language teachers	-9.66	-0.35	.000*
Assessment in language pedagogy	LTA developers & LTA researchers	-6.55	-0.36	.000*
	LTA developers & language teachers	-6.62	-0.23	.000*
	LTA researchers & language teachers	-2.44	-0.09	.015
Assessment policy and local practices	LTA developers & LTA researchers	-3.27	-0.18	.001*
	LTA developers & language teachers	-6.98	-0.24	.000*
	LTA researchers & language teachers	-8.75	-0.31	.000*
Personal beliefs and attitudes	LTA developers & LTA researchers	-0.28	-0.02	.783
	LTA developers & language teachers	-6.02	-0.21	.000*
	LTA researchers & language teachers	-5.79	-0.21	.000*
Statistical and research methods	LTA developers & LTA researchers	-1.68	-0.09	.092
	LTA developers & language teachers	-13.22	-0.46	.000*
	LTA researchers & language teachers	-12.66	-0.45	.000*
Assessment principles and interpretation	LTA developers & LTA researchers	-0.25	-0.01	.806
	LTA developers & language teachers	-11.66	-0.40	.000*
	LTA researchers & language teachers	-10.37	-0.37	.000*
Language structure, use and development	LTA developers & LTA researchers	-0.59	-0.03	.555
	LTA developers & language teachers	-2.98	-0.10	.000*
	LTA researchers & language teachers	-3.29	-0.12	.001*
Scoring and rating	LTA developers & LTA researchers	-0.13	-0.01	.133
	LTA developers & language teachers	-9.75	-0.34	.000*
	LTA researchers & language teachers	-6.72	-0.24	.000*

Table 8 shows that LTA developers’ and LTA researchers’ responses were almost indistinguishable across the dimensions, with the exception of *Assessment in language pedagogy* and *Assessment policy and local practices* where there was a medium and small effect (respectively) in the direction of LTA researchers. On the remaining six dimensions, however, both developers and researchers were observed to rate their needs significantly higher than did language teachers, with the largest effect sizes observed on *Statistical and research methods*, *Developing and administering language assessments*, and *Assessment principles and interpretation* (all in the medium-strength range according to Cohen [1988]). These findings are, for the most part, expected, as LTA test developers and researchers require a high level of expertise in developing and evaluating language assessments, and a deep understanding of the constructs which underlie them and the principles which guide assessment practice. However, it is less encouraging that teachers would rate their needs in *Scoring and rating* lower than the developer and research groups given that scoring and rating may fall directly under the professional responsibility of language educators. This finding is, however, in line with Vogt and Tsagari's (2014) study in which teachers did not consider "giving grades" a particularly important feature of teacher training.

Assessment in language pedagogy and *Assessment policy and local practices* provide interesting counterpoints to the pattern. In the former, there was no statistically significant difference between the ratings of the LTA researchers and language teachers, though both groups rated their needs here higher than did the LTA developers. In the latter, a hierarchy was observed in self-perceived needs, with LTA researchers the highest, followed by LTA developers, and then language teachers. LTA developers' responses here are less surprising considering that most respondents in this role who responded to the survey would be likely to work for larger-scale assessment organizations or exam boards where more local classroom assessment issues and policy requirements are considered beyond the scope of their day-to-day professional concern. For language teachers, the finding with respect to *Assessment policy and local practices* may indicate that policy- and regulatory-issues are viewed as a concern of the management or leadership teams rather than of teachers themselves. However, some items within this dimension – e.g., how to determine if the results from a language assessment are relevant to the local context – appear to be crucially important to classroom practice.

Due to the evolution of dimensions (both in label and composition) it is difficult to directly compare these empirical profiles to those suggested by Taylor (2013) (see Figure 1). Nonetheless, some useful observations can be made in comparison with the profiles for (a), (b) and (d). For example, Taylor's supposition that LTA researchers would need an expansive and balanced type of LAL was supported empirically. While the survey responses did not reach the extreme levels of Taylor's profiles, this is likely an artefact of some respondents avoiding the very highest category in the survey. Taylor's profile for test developers was similarly supported, though with a slightly more rounded profile emerging from the survey (particularly with respect to the dimensions related to personal beliefs and local practices). The lower rating for language pedagogy, however, was upheld in the survey results. Finally, the teacher profile matched Taylor's predictions in the sense that it was reduced in scope compared to the other two groups. However, while the Taylor profile speculates that understanding of local practices and sociocultural values are relatively important (Level 3) for language teachers, this was not borne out in the empirical data. There is also a discrepancy with the newly collated category “assessment policy and local practices”, which teachers only thought they needed to be “moderately knowledgeable” about (in contrast to Taylor’s speculation of a higher need for the original dimensions). Taylor’s hypothesis that knowledge about the relation of assessment to language pedagogy would be more important for this group than all other aspects of LAL, could not be confirmed (as discussed above). In contrast, knowledge of (language) theory and understanding of assessment principles were rated as more important for teachers by teachers themselves than Taylor’s profiles would have led us to expect.

Conclusion

The present study has aimed to provide a clearer understanding of what developmental components LAL comprises. It has done so by attempting to validate the LAL profile model suggested by Taylor (2013) from a synthesis of recent literature on LAL frameworks, through the development of a large-scale online survey that has invited different stakeholder groups to indicate the level of LAL they think members of their group are required to have in order to perform their assessment-related tasks well. Through this, a survey tool has been carefully developed with expert feedback, which has been empirically reduced down to a feasible instrument through EFA on a large-scale set of responses (even though we do recommend retaining the full set of 71 items if the survey is to be used for diagnostic purposes, e.g., for performing a needs analysis where knowledge/skills of specific items are of interest). This now allows us to both describe the component structure of LAL more systematically as well as investigate the LAL needs profiles that different stakeholder groups identify for themselves. The results from the study have suggested that there may be nine distinct components of LAL, which are largely in line with Taylor’s (2013) hypothesised components, but with some key distinctions or expansions: Developing and administering language assessments, Assessment in language pedagogy, Assessment policy and local practices, Personal beliefs and attitudes, Statistical and research methods, Assessment principles and interpretation, Language structure, use and development, Washback and preparation, and Scoring and rating. Using this structure, it was possible to gauge the LAL needs of different stakeholder groups as perceived by themselves in a developmental profile model that showed differences between the LAL requirements of various groups, and could also be employed to describe the needs and wants across different geographical contexts.

The study and approach taken has limitations. We have acknowledged that the population sample in the data is not fully balanced across stakeholder groups, and may be skewed due to the channels used to distribute the survey. While the number of teacher respondents is sizeable, there were few responses from parents, students and policy-makers. This is likely to be attributable both to the lengthy and somewhat technical nature of the instrument itself, and because these groups are in any case challenging for language testing researchers to reach and involve in research (e.g. Malone, 2016). The fact that the instrument was only made available in English for the purpose of this research may have functioned as another constraint. Further research that will recruit new respondents may also usefully to employ CFA to corroborate the factor structure of this instrument.

A further limitation is that we cannot be certain that respondents interpreted the items in the same way (although this is a limitation of any survey instrument, and we aimed to minimize this limitation through extensive pre-testing). Finally, although the development process involved a wide range of input, there may be elements of LAL that were not adequately captured in the survey items. However, these limitations create opportunities for further research: triangulating the quantitative data with qualitative methods (think-alouds, interviews) to establish how items are understood, and comparing the range of dimensions with LAL needs elicited in more exploratory, contextualized studies. Such research could then particularly strive to engage non-experts. Further qualitative research on the use of the questionnaire with ‘lay’ groups with a view to creating more targeted and user-friendly items would need to address this.

The Language Assessment Literacy Survey will remain an open-access tool, available for the purposes of needs analyses, self-assessment, and reflective practice across different contexts. It is our hope that further use of the questionnaire will help to build a large dataset of LAL needs across different geographical contexts which might be compared and contrasted for the purposes of a better understanding of LAL generally. The survey is already in use in Mongolia, Uzbekistan, Turkey, Vietnam, Brazil, USA, China, and the Netherlands, and has been successfully translated and adapted to local contexts. At the same time, we recognize the limitations of the “broad-brush” understandings that any survey can generate, and we therefore hope other researchers may make use of the survey in mixed-methods designs to provide more contextualized qualitative perspectives on these issues.

References

Baker, B. (2016). Language assessment literacy as professional competence: The case of Canadian admissions decision makers. *Canadian Journal of Applied Linguistics, 19*(1), 63-83.

Baker, B. A., & Riches, C. (2017). The development of EFL examinations in Haiti: Collaboration and language assessment literacy development. *Language Testing*, 1–25.

Brindley, G. (2001). Language assessment and professional development. In C. Elder, A. Brown, E. Grove, K. Hall, N. Iwashita, T. Lumley et al. (Eds.), *Experimenting with Uncertainty: Essays in Honour of Alan Davies* (pp 126–136). Cambridge: Cambridge University Press.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Lawrence Earlbaum Associates.

Davies, A. (2008). Textbook trends in teaching language testing. *Language Testing*, *25*(3), 327–347.

Deygers, B., & Malone, M. E. (2019). Language assessment literacy in university admission policies, or the dialogue that isn't. *Language Testing*, 0265532219826390.

Dörnyei, Z., & Taguchi, T. (2010). *Questionnaires in second language research: Construction, administration, and processing*. New York: Routledge.

Dreyfus, S., & Dreyfus, H. (1980). *A Five-Stage Model of the Mental Activities Involved in Directed Skill Acquisition*. http://www.dtic.mil/cgi-bin/GetTRDoc?AD=ADA084551&Location=U2&doc=GetTRDoc.pdf

van der Eijk, C., & Rose, J. (2015). Risky business: factor analysis of survey data - assessing the probability of incorrect dimensionalisation. *PLoS One*, *20*(3), 1–31.

Fulcher, G. (2012). Assessment Literacy for the Language Classroom. *Language Assessment Quarterly*, *9*(2), 113–132. doi:10.1080/15434303.2011.642041

Harding, L., & Kremmel, B. (2016). Teacher assessment literacy and professional development. In D. Tsagari & J. Banerjee (Hrsg.), *Handbook of Second Language Assessment* (S. 413–428). Berlin: De Gruyter.

Harding, L., & Kremmel, B. (in preparation). The language assessment literacy profiles of different stakeholder groups in different contexts: Needs, lacks and wants from a large-scale survey.

Inbar-Lourie, O. (2008). Constructing a language assessment knowledge base: A focus on language assessment courses. *Language Testing*, *25*(3), 385–402. doi:10.1177/0265532208090158

Lam, R. (2015). Language assessment training in Hong Kong: Implications for language assessment literacy. *Language Testing, 32*(2), 169-197.

Loewen, S., & Gonulal, T. (2015). Exploratory factor analysis and principal components analysis. In Plonsky, L. (Ed), *Advancing quantitative methods in second language research* (pp. 182-212). New York: Routledge

Malone, M. E. (2016). *Expanding understanding of language assessment literacy: Including students.* Invited Plenary at the Language Assessment Literacy Symposium, University of Lancaster, Lancaster, UK.

Mertler, C. A. (2009). Teachers’ assessment knowledge and their perceptions of the impact of classroom assessment professional development. *Improving Schools, 12*(2), 101–113. https://doi.org/10.1177/1365480209105575

Mertler, C.A. & Campbell, C.S. (2005) Measuring teachers' knowledge and application of classroom assessment concepts: development of the Assessment Literacy Inventory. Paper presented at the annual meeting of the American Educational Research Association, Montreal, Quebec, Canada, April.

Mokken, R. J. (1971). *A theory and procedure of scale analysis*. The Hague/Berlin: Mouton/De Gruyter.

O’Loughlin, K. (2013). Developing the assessment literacy of university proficiency test users. *Language Testing*, *30*(3), 363–380.

Pill, J., & Harding, L. (2013). Defining the language assessment literacy gap: Evidence from a parliamentary inquiry. *Language Testing*, *30*(3), 381–402.

Plake, B. S., Impara, J. C. and Fager, J. J. (1993), Assessment Competencies of Teachers: A National Survey. Educational Measurement: Issues and Practice, 12: 10-12. doi:10.1111/j.1745-3992.1993.tb00548.x

Stiggins, R. J. (1991). Assessment literacy. *Phi Delta Kappan*, *72*(7), 534–39.

Taylor, L. (2009). Developing Assessment Literacy. *Annual Review of Applied Linguistics, 29*, 21–36. doi:10.1017/S0267190509090035

Taylor, L. (2013). Communicating the theory, practice and principles of language testing to test stakeholders: Some reflections. *Language Testing, 30*(3), 403–412. doi:10.1177/0265532213480338

Vogt, K., & Tsagari, D. (2014). Assessment Literacy of Foreign Language Teachers: Findings of a European Study. *Language Assessment Quarterly, 11*(4), 374-402.

Yan, X., Zhang, C. & Fan, J.J. (2018). “Assessment knowledge is important, but…”: How contextual and experiential factors mediate assessment practice and training needs of language teachers." *System*, *74*, 158-168.

Appendix 1 – Administered survey: instructions, items and hypothesized codes

The full set of items and instructions used in the main administration of the survey (launched May 2017) is shown below together with codes showing dimensions hypothesized prior to administration (based on an extension of Taylor’s [2013] profiles). We encourage others to make use of these items with due attribution to the current paper.

How knowledgable do people in your chosen group/profession *need to be* about each aspect of language assessment below? Please respond according to the following scale:

Not knowledgeable at all / slightly knowledgeable / moderately knowledgeable / very knowledgeable / extremely knowledgeable

1) how to use assessments to inform learning or teaching goals (LangP)
2) how to use assessments to evaluate progress in language learning (LangP)
3) how to use assessments to evaluate achievement in language learning (LangP)
4) how to use assessments to evaluate language programs (LangP)
5) how to use assessments to diagnose learners’ strengths and weaknesses (LangP)
6) how to use assessments to motivate student learning (LangP)
7) how to use self-assessment (LangP)
8) how to use peer-assessment (LangP)
9) how to interpret measurement error (SDM)
10) how to interpret what a particular score says about an individual’s language ability (SDM)

11) how to determine if a language assessment aligns with a local system of accreditation (LocP)
12) how to determine if a language assessment aligns with a local educational system (LocP)
13) how to determine if the content of a language assessment is culturally appropriate (LocP)
14) how to determine if the results from a language assessment are relevant to the local context (LocP)

15) how to communicate assessment results and decisions to teachers (SDM)
16) how to communicate assessment results and decisions to students or parents (SDM)

17) how to train others about language assessment (LangP)
18) how to recognize when an assessment is being used inappropriately (ISV)
19) how to prepare learners to take language assessments (LangP)
20) how to find information to help in interpreting test results (SDM)
21) how to give useful feedback on the basis of an assessment (LangP)

22) how assessments can be used to enforce social policies (e.g., immigration) (ISV)
23) how assessments can influence teaching and learning in the classroom (LangP)
24) how assessments can influence teaching and learning materials (LangP)
25) how assessments can influence the design of a language course or curriculum (LangP)

26) how language skills develop (e.g., reading, listening, writing, speaking) (KT)
27) how foreign/second languages are learned (KT)
28) how language is used in society (KT)
29) how social values can influence language assessment design and use (ISV)
30) how pass-fail marks / cut-scores are set (SDM)

31) the concept of reliability (how accurate or consistent an assessment is) (PC)
32) the concept of validity (how well an assessment measures what it claims to measure) (PC)
33) the structure of language (KT)
34) the advantages and disadvantages of standardized testing (ISV)
35) the history of language assessment (ISV)
36) the philosophy behind the design of a relevant language assessment (ISV)
37) the impact language assessments can have on society (ISV)
38) the relevant legal regulations for assessment in the local area (LocP)
39) the assessment traditions in a local context (LocP)
40) the specialist terminology related to language assessment (PC)

41) different language proficiency frameworks (e.g., the Common European Framework of Reference [CEFR]) (KT)
42) different stages of language proficiency (KT)
43) different types of purposes for language assessment purposes (e.g., proficiency, achievement, diagnostic) (PC)
44) different forms of alternative assessments (e.g., portfolio assessment) (PC)

45) one's own beliefs/attitudes towards language assessment (PBA)
46) how one's own beliefs/attitudes might influence one's assessment practices (PBA)
47) how one's own beliefs/attitudes may conflict with those of other groups involved in assessment (PBA)
48) how one's own knowledge of language assessment might be further developed (PBA)

How skilled do people in your chosen group/profession *__need to be__* in each aspect of language assessment below? Please respond according to the following scale:

Not skilled at all / slightly skilled / moderately skilled / very skilled / extremely skilled

49) using statistics to analyse the difficulty of individual items (questions) or tasks (TS-C)
50) using statistics to analyse overall scores on a particular assessment (TS-C)
51) using statistics to analyse the quality of individual items/tasks (TS-C)
52) using techniques other than statistics (e.g., questionnaires, interviews, analysis of language) to get information about the quality of a language assessment (TS-C)
53) using rating scales to score speaking or writing performances (TS-B)
54) using specifications to develop items and tasks (TS-A)

55) scoring closed-response questions (e.g. Multiple Choice Questions) (TS-B)
56) scoring open-ended questions (e.g. short answer questions) (TS-B)

57) developing portfolio-based assessments (TS-A)
58) developing specifications (overall plans) for language assessments (TS-A)

59) selecting appropriate rating scales (rubrics) (TS-A)
60) selecting appropriate items or tasks for a particular assessment purpose (TS-A)

61) training others to use rating scales (rubrics) appropriately (TS-B)
62) training others to write good quality items (questions) or tasks for language assessments (TS-A)

63) writing good quality items (questions) or tasks for language assessments (TS-A)
64) aligning tests to proficiency frameworks (e.g., the Common European Framework of Reference) (TS-C)

65) determining pass-fail marks / cut-scores (TS-C)
66) identifying assessment bias (TS-C)
67) accommodating candidates with disabilities or other learning impairments (TS-B)
68) designing scoring keys and rating scales (rubrics) for assessment tasks (TS-A)
69) making decisions about what aspects of language to assess (TS-A)
70) piloting/trying-out assessments before their administration (TS-A)
71) selecting appropriate ready-made assessments (TS-A)

Items per hypothesized dimension:

Knowledge of theory (KT) = 6
Principles and concepts (PC) = 5
Language pedagogy (LangP) = 14
Impact and social values (ISV) = 7
Local practices (LocP) = 6
Personal beliefs/attitudes (PBA) = 4
Scores and decision-making (SDM) = 6
Technical skills (A) – Constructing language assessments (TS-A) = 11
Technical skills (B) – Administering/scoring language assessments (TS-B) = 5
Technical skills (C) – Evaluating language assessments (TS-C) = 7

Total = 71

Appendix 2 – List of removed items

No.	Item
2)	how to use assessments to evaluate progress in language learning
4)	how to use assessments to evaluate language programs
9)	how to interpret measurement error
15)	how to communicate assessment results and decisions to teachers
16)	how to communicate assessment results and decisions to students or parents
20)	how to find information to help in interpreting test results
30)	how pass-fail marks / cut-scores are set
34)	the advantages and disadvantages of standardized testing
35)	the history of language assessment
36)	the philosophy behind the design of a relevant language assessment
37)	the impact language assessments can have on society
40)	the specialist terminology related to language assessment
41)	different language proficiency frameworks (e.g., the Common European Framework of Reference [CEFR])
42)	different stages of language proficiency
44)	different forms of alternative assessments (e.g., portfolio assessment)
54)	using specifications to develop items and tasks
57)	developing portfolio-based assessments
71)	selecting appropriate ready-made assessments

Appendix 3: Rotated pattern matrix with factor loadings

	Factor								
	1	2	3	4	5	6	7	8	9
62) training others to write good quality items (questions) or tasks for language assessments	.801								
68) designing scoring keys and rating scales (rubrics) for assessment tasks	.758								
61) training others to use rating scales (rubrics) appropriately	.730								
63) writing good quality items (questions) or tasks for language assessments	.717								
64) aligning tests to proficiency frameworks (e.g., the Common European Framework of Reference [CEFR], American Council on the Teaching of Foreign Languages [ACTFL])	.654								
66) identifying assessment bias	.652								
70) piloting/trying-out assessments before their administration	.598								
69) making decisions about what aspects of language to assess	.587								
65) determining pass-fail marks or cut-scores	.585								
60) selecting appropriate items or tasks for a particular assessment purpose	.519								
67) accommodating candidates with disabilities or other learning impairments	.518								
58) developing specifications (overall plans) for language assessments	.478								
59) selecting appropriate rating scales (rubrics)	.476								
17) how to train others about language assessment	.445								
8) how to use peer-assessment		.862							
7) how to use self-assessment		.857							
6) how to use assessments to motivate student learning		.590							

5) how to use assessments to diagnose learners' strengths and weaknesses	.454							
1) how to use assessments to guide learning or teaching goals	.449							
21) how to give useful feedback on the basis of an assessment	.362							
12) how to determine if a language assessment aligns with a local educational system		.838						
11) how to determine if a language assessment aligns with a local system of accreditation		.796						
38) the relevant legal regulations for assessment in your local area		.572						
14) how to determine if the results from a language assessment are relevant to the local context		.569						
39) the assessment traditions in your local context		.490						
22) how assessments can be used to enforce social policies (e.g., immigration, citizenship)		.430						
46) how your own beliefs/attitudes might influence one's assessment practices			-.967					
47) how your own beliefs/attitudes may conflict with those of other groups involved in assessment			-.867					
45) your own beliefs/attitudes towards language assessment			-.825					
48) how your own knowledge of language assessment might be further developed			-.567					
50) using statistics to analyse overall scores on a particular assessment				.889				
49) using statistics to analyse the difficulty of individual items (questions) or tasks				.883				
51) using statistics to analyse the quality of individual items (questions)/tasks				.882				
52) using techniques other than statistics (e.g., questionnaires, interviews, analysis of language) to get information about the quality of a language assessment				.531				
32) the concept of validity (how well an assessment measures what it claims to measure)					.666			
31) the concept of reliability (how accurate or consistent an assessment is)					.618			

3) how to use assessments to evaluate achievement in language learning					.380			
10) how to interpret what a particular score says about an individual's language ability					.374			
28) how language is used in society						.862		
27) how foreign/second languages are learned						.697		
26) how language skills develop (e.g., reading, listening, writing, speaking)						.590		
29) how social values can influence language assessment design and use						.441		
33) the structure of language						.410		
24) how assessments can influence teaching and learning materials							-.828	
25) how assessments can influence teaching and learning in the classroom							-.732	
23) how assessments can influence the design of a language course or curriculum							-.603	
19) how to prepare learners to take language assessments							-.345	
56) scoring open-ended questions (e.g. short answer questions)								-.504
55) scoring closed-response questions (e.g. Multiple Choice Questions)								-.437
53) using rating scales to score speaking or writing performances								-.375

Extraction Method: Principal Axis Factoring.
Rotation Method: Oblimin with Kaiser Normalization.[a]

a. Rotation converged in 12 iterations.

Appendix 4 – Descriptive statistics of LAL needs for three key stakeholder groups

	LTA developers (*n*=198)		**LTA researchers (*n*=138)**		**Language teachers (*n*=645)***	
	M	**SD**	**M**	**SD**	**M**	**SD**
Developing and administering language assessments	3.35	.59	3.28	.60	2.53	.87
Assessment in language pedagogy	2.53	.83	3.12	.70	2.96	.72
Assessment policy and local practices	2.75	.77	3.01	.82	2.28	.86
Personal beliefs and attitudes	3.21	.85	3.28	.74	2.83	.89
tatistical and research methods	3.25	.80	3.38	.74	2.10	1.03
Assessment principles and interpretation	3.60	.52	3.63	.49	2.94	.79
Language structure, use and development	3.19	.70	3.25	.61	3.02	.73
Washback and preparation	2.85	.82	3.04	.74	3.01	.79
Scoring and rating	3.45	.68	3.31	.79	2.83	.83

* Note, for the Language teachers group: *n*=644 for *Personal beliefs and attitudes* and *Assessment principles and interpretation*; *n*=643 for *Statistical and research methods* and *Scoring and rating*

[i] Of these, 91 surveys were removed because confidence was below 50%. The remainder were removed because surveys were incomplete. Of the 91 low-confidence responses, the following proportions of role/profession were recorded: language testing/assessment developers (18%); language testing/assessment researchers (11%); language teachers (64%); parents (2%); policy-maker (1%); test-score user (1%); test-taker (3%)