Teacher Assessment Literacy Scale: Design and Validation

Kay Cheng Soh and Limei Zhang

Abstract With the increased use of language assessment, language teachers are required to develop assessment literacy and take the testing and assessment responsibilities. This article reports the processes and outcomes of the development and validation of the Teacher Assessment Literacy Scale for use with Singapore's Chinese Language teachers.

In the past, patients were passive recipients of medical treatments. The present-day patients are involved in the healing process; they are informed and they are engaged. Analogously, in the past, student assessment tools were crafted by test specialists while teachers were passive users; this is true at least in the American context where standardized tests are the regular fixture of the school. Nowadays, with the emphasis on assessment *for* learning (or *formative* assessment) in contrast with assessment *of* learning (or *summative* assessment), teachers, in America and elsewhere, are expected to use assessment in a more engaged manner to help students learn. Teachers are therefore expected to use test information not only for assessment *of* learning but also, perhaps more importantly, assessment *for* learning. This shift all the more underlines the importance of teachers' assessment literacy if they were to complete this crucial aspect of their job with professionalism. Popham (2006) brought up his very apt analogy between educational and healthcare professions where proper use of test information is crucial. Not only do teachers need assessment literacy but everyone else who has an interest in education and *everyone* includes teachers, school leaders, policy-makers, and parents.

Due to the change in the emphasis on formative assessment and its contribution to learning (Fulcher 2012), the notion of assessment literacy has changed contingently. Traditionally, assessment emphasizes objectivity and accuracy (Spolsky

K. C. Soh (✉) · L. Zhang
Nanyang Technological University, Singapore, Singapore
e-mail: Kaycheng.soh@sccl.sg

L. Zhang
e-mail: Limei.zhang@sccl.sg

K. C. Soh (ed.), *Teaching Chinese Language in Singapore*,
https://doi.org/10.1007/978-981-10-8860-5_13

1978, 1995), due to the influence of the psychometric and positivistic paradigms, and testing activities normally are carried out at the end of learning periods (Gipps 1994; Wolf et al. 1991). In that context, only measurement specialists were expected to have specialized knowledge of test development, score interpretation, and theoretical concepts of measurement. In contrast, assessment is now perceived as an integral part of teaching and learning to provide timely information as feedback to guide further instruction and learning. This requires teachers to design assessment, make use of test results to promote teaching and learning, and be aware of inherent technical problems and limitation of educational measurement (Fulcher 2012; Malone 2008). Thus, it is important that teachers have sufficient practical skills as well as theoretical understanding.

Assessment Literacy Measures

Over the years, efforts have been made to measure teacher assessment literacy. Gotch and French (2014) systematically reviewed teacher assessment literacy measures within the context of contemporary teacher evaluation policy. The authors collected objective tests of assessment knowledge, teacher self-reports, and rubrics to evaluate teachers' work in assessment literacy studies from 1991 to 2012. They then evaluated the psychometric work from these measures against a set of claims related to score interpretation and use. Across the 36 measures reviewed, they found weak support for these claims. This highlights the need for increased work on assessment literacy measures in the educational measurement field.

DeLuca et al. (2016) emphasized that assessment literacy is a core professional requirement across educational systems and that measuring and supporting teachers' assessment literacy have been a primary focus over the past two decades. At present, according to the authors, there is a multitude of assessment standards across the world and numerous assessment literacy measures representing different conceptions of assessment literacy. The authors analyzed assessment literacy standards from five English-speaking countries (i.e., Australia, Canada, New Zealand, the UK, and the USA) and Europe to understand shifts in the assessment developed after 1990. Through a thematic analysis of 15 assessment standards and an examination of eight assessment literacy measures, the authors noticed shifts in standards over time though the majority of the measures continue being based on early conceptions of assessment literacy.

Stiggins (1991) first coined the term *assessment literacy* to refer to teachers' understanding of the differences between sound and unsound assessment procedures and the use of assessment outcomes. Teachers who are assessment literates should have a clear understanding of the purposes and targets of assessment, the competence in choosing appropriate assessment procedures, the capability of conducting assessment effectively and of avoiding pitfalls in the process of assessment practices and interpretation of results.

This sounds simple but can be a tall order in actuality. For example, the by now classic textbook of educational measurement by Linn and Miller (2005) has altogether 19 chapters in three parts. The five chapters in Part I cover such topics on the role of assessment, instructional goals of assessment, concepts of reliability and validity, and issues and trends. These may not be of immediate relevance to the classroom teachers' work but provide necessary conceptual backgrounds for teachers to be informed assessors. Part II has ten chapters of a technical or procedural nature, which equip teachers with the necessary practical skills in test design using a wide range of item formats. The ending Part III has four chapters, dealing with selecting and using published tests as well as the interpretation of scores involving basic statistical concepts. The three parts that made up the essential domains of assessment literacy expected of classroom teachers are typical of many education measurement texts supporting teacher training programs.

According to Popham (2009), increasing numbers of professional development programs have dealt with assessment literacy for teachers and administrators. Popham then asked the question of whether assessment literacy is merely a fashionable focus or whether it should be regarded as a significant area of professional development for years to come. Popham first divided educators' measurement-related concerns into either classroom assessments or accountability assessments and then argued that educators' inadequate knowledge about either of these can cripple the quality of education. He concluded that assessment literacy is a *condicio sine qua non* for today's competent educator and must be a pivotal content area for current and future staff development.

The above review of the pertinent literature on assessment literacy and its measurement has implications for the present study. First, in recent years, the Singapore Ministry of Education has launched the initiatives emphasizing higher-order thinking skills and deep understanding in teaching, such as 'Teach Less, Learn More' (TLLM) and 'Thinking Schools, Learning Nations' (TSLN). Consequently, school teachers are required to make changes to their assessment practice and to equip themselves with sufficient assessment literacy. In spite of the importance of assessment literacy, few studies have been conducted to examine their assessment knowledge and skills. Among the very few local studies, Koh (2011) investigated the effects of professional development on Primary 4 and 5 teachers of English, Science, and Mathematics. She found that ongoing professional development of assessment literacy is especially effective in improving teachers' assessment literacy, when compared with teachers who participated in workshops training them to design assessment rubrics. The findings suggest that to successfully develop teachers' assessment literacy, the training needs be broad enough in the topics covered, and the training has to be extended over a reasonable period of time.

In a more recent study, Zhang & Chin (under review) examined the learning needs in language assessment among 103 primary school Chinese Language teachers using an assessment literacy survey developed by Fulcher (2012). Results provide an understanding of teachers' interest and knowledge in test design and development, large-scale testing, classroom testing, and test validity and reliability. With these very limited number of studies in the Singapore context, there is a need for more studies to be carried out for a better understanding of Singapore school teachers' assessment literacy. For carrying out such studies, it is necessary to develop an assessment literacy scale which is broad enough and yet concise to measure the teachers' assessment competence properly and accurately.

Secondly, in systems like that of the USA where standardized tests are designed by test specialists through a long and arduous process of test development, applying sophisticated psychometric concepts and principles (with regular intermittent revisions), it is reasonable to assume that the resultant assessment tools made available for teachers are of a high psychometric quality. In such a case, the most critical aspect of assessment literacy that teachers need is the ability to properly interpret the results they obtain through the tests. Measurement knowledge beyond this is good to have but not really needed. However, in a system like that of Singapore where standardized tests are *not* an omnipresent fixture, teacher-made tests are almost the only assessment tool available. This indicates the teachers' need for assessment literacy of a much broader range, going beyond the interpretation of test results. Therefore, this study aims to develop an instrument for assessment literacy to measure teachers' assessment literacy in the Singapore context.

Thirdly, previous studies have provided the framework for the present writers to follow in designing an assessment literacy scale. As one of the most influential studies in language assessment literacy, Fulcher (2012) has expanded the definition of assessment literacy. According to him, assessment literacy comprises three levels' knowledge:

- Level 1 concerns the knowledge, skills, and abilities in the practice of assessment, especially in terms of test design. Specifically, this type of knowledge includes how to decide what to test, writing test items and tasks, developing writing test specifications, and developing rating scales.
- Level 2 refers to the processes, principles, and concepts of assessment, which are more relevant to quality standards and research. This type of knowledge includes validity, reliability, fairness, accommodation, washback/consequences as well as ethics and justice of assessment.
- Level 3 is about historical, social, political, philosophical, and ethical frameworks of assessment, which is concerned with such issues as the historical, social, and political as well as the philosophical and ethical bases for assessment practice.

Following Fulcher's (2012) framework, we aim to measure teachers' assessment knowledge of two aspects, i.e., (1) the knowledge, skills, and abilities in assessment practice as well as the fundamental principles and (2) concepts of language assessment. This does not mean that we did not value the third domain (Level 3) but that we considered this as not being so urgently needed by teachers in Singapore and as not being so critical to their day-to-day use of assessment in the classroom context.

Method

Design

In the development of the *Teacher Assessment Literacy Scale*, consultation was made to two classics of educational measurement (Hopkins 1998; Linn and Miller 2005). The first decision to be made was to identify and delimit the domains to be covered in the scale, and it was decided that four key domains needed to be represented in the new measure:

1. Understanding of the nature and functions of assessment.
2. Practical skills to design and use a variety of item formats to meet the instructional needs.
3. Ability to properly interpret these to inform further teaching and guide student learning.
4. Ability to evaluate the qualities of the test results, involving basic knowledge of statistics.

Against the background of the above considerations, it was decided that ten items be written for each domain as a sample representing the possible items of the domain. Four domains having been delimited, the whole scale thus comprises 40 items. It was further decided that the items would take the form of four-option multiple choice to ensure objectivity in scoring and keep the testing time within a reasonable limit of about 30 min.

Due to space constrain, the actual items are not presented in this article. However, interested researchers may contact the first author for a copy of the items via sohkaycheng@hotmail.com.

Trial Sample

The scale thus crafted was then administered to 323 Chinese Language teachers, 170 from primary schools, and 153 from secondary schools and junior colleges. There is a female preponderance of 83%. Of the teachers, 52% have more than ten years of teaching experience. In terms of qualification, 93% hold a university degree and 95% have completed professional training. However, only 48% reported that they had elected to study assessment in their pre-service training, and 78% acknowledged that they felt the need for more training in assessment.

Teachers attended various in-service courses at the Singapore Centre for Chinese Language from January to March 2015. The participants can be considered mature in the teaching profession as more than half of them having ten or more years of teaching experience. Moreover, the female preponderance is typical of the teaching profession in Singapore. Thus, bearing in mind some limitations in these regards, the participating Chinese Language teachers can be deemed sufficiently representative of Chinese Language teachers in Singapore.

Analysis

Confirmatory factor analysis was performed to examine whether the collected data support the proposed model of four specified dimensions. Next, the classical item analysis was performed to obtain item difficulty (*p*) and item discrimination (*r*). Then, the Rasch analysis was conducted to estimate item locations to indicate item difficulty within the context of the whole set of items analyzed, with a positive index indicating the *difficulty to answer correctly* and vice versa.

Results

Results of statistical analysis are highlighted below, and details can be found in Soh and Zhang (2017).

Descriptive Statistics

As Table 1 shows, the means for the subscales vary between 5.52 and 2.97, out of 10. The highest mean is for the first subscale (nature and function of assessment) while the lowest is for the fourth subscale (concepts of reliability, validity, etc). Generally, the means show that teachers were able to answer correctly about half of the 30 questions in subscales 1–3, but they were able to answer correctly only about three of the 10 questions in subscale 4. If a criterion-referenced approach requiring 90% of the teachers to be able to answer correctly 90% of the questions, thus expecting approximately 80% of correct responses, the results obtained are far from being satisfactory, being 45% on average.

Table 1 Descriptive statistics

Subscale	Mean	Standard deviation
Nature and function of assessment	5.53	1.23
Design and use of test items	4.87	1.38
Interpretation of test results	4.50	1.58
Concepts of reliability, validity, etc.	2.97	1.50

Confirmatory Factor Analysis

When confirmatory factor analysis was run, the results show that the incremental fit index CFI = 1.00 is greater than 0.95, while the absolute fit index RMSEA = 0.001 is less than 0.06. The RMSEA 95% confidence interval is narrow. $X^2/df = 0.57$ is less than 3.00.

Table 2 shows the path coefficients of assessment literacy range from 0.21 (concepts of reliability, validity, etc) to 0.49 (nature and function of assessment), all significant at the 0.05 level, and the average is 0.39, indicating that the latent variable is well-defined by the four variables. However, of the four path coefficients, those for the first three subscales are sizable, varying from 0.41 to 0.49, but that for the fourth subscale (statistical and measurement concepts) is rather low at 0.21, indicating that the other three subscales are better measures of assessment literacy.

Classical Item Analysis

Subtest 1: Nature and Functions of Assessment

Subtest 1 deals with understanding the functions assessment has in teaching and learning, and concepts related to the norm- and criterion-referenced interpretation of test scores. For this subtest, the facility indices (*p*) vary from a very low 0.07 to a very high 0.94, with a mean of 0.55 and a median of 0.56. In short, the items of Subtest 1 vary widely in difficulty although the average facility suggests that this subtest as a whole is moderately difficult. At the same time, the discrimination indices (*r*) vary from 0.13 to 0.33, with a mean discrimination of 0.23 and a median of 0.22. These figures indicate that the items have a low but acceptable discriminatory power.

Subtest 2: Design and Use of Test Items

The items of Subtest 2 deal with the understanding of the suitability of various item formats and their appropriate uses. The *p*'s vary from a low 0.13 to a very high 0.96, with a mean of 0.49 and a median of 0.44. In short, these items vary widely in

Table 2 Path coefficients

Subscale	Path coefficient	Error
Nature and function of assessment	0.49	0.24
Design and use of test items	0.41	0.17
Interpretation of test results	0.47	0.22
Concepts of reliability, validity, etc.	0.21	0.64

difficulty although the mean suggests that this subtest is moderately difficult as a whole. At the same time, the *r*'s vary from 0.11 to 0.30, with a mean of 0.21 and a median of 0.22. These results indicate that the items have a low but.

Subtest 3: Interpretation of Test Results
The items of Subtest 3 pertain to knowledge of item indices and meanings of test scores. The *p'*s vary from a very low 0.03 to a high 0.78, with a mean of 0.45 (median 0.51). These figures indicate that the items vary widely in difficulty although the mean suggests that this subtest is of moderate difficulty. At the same time, the *r*'s vary from 0.05 to 0.47, with a mean of 0.24 (median 0.23). These results indicate that the subtest as a whole has acceptable discrimination.

Subtest 4: Concepts of Reliability, Validity, and Basic Statistics
Subtest 4 deals with abstract concepts of test score qualities and knowledge of simple statistics essential to understand test results. The *p*'s vary from a low 0.11 to a high 0.64, with a mean of 0.30 (median 0.25). These figures indicate that the items are difficult ones when compared with those of the other three subtests. The *r*'s vary from 0.05 to 0.36, with a mean of 0.19 (median 0.17). These results indicate that the subtest as a whole has low discrimination.
By way of summary, the 40 items generally have an acceptable level of facility for the teachers involved in this study, although the facilities and discrimination vary widely, reflecting probably the wide heterogeneity of the teachers' assessment literacy. Moreover, the items tend to be have low discrimination power, partly due to the constrains of low facilities of some items. These findings could at least partly account for the fact that the teachers taking part in this study have discernible deficits in their assessment literacy, with an overall mean of 18 for the 40 items (i.e., 45%).

Rasch Analysis

For the Rasch analysis (Rasch 1993) performed on the 40 items, item estimates vary from −3.664 to +3.245, with a mean of 0.000 (median 0.260). These show that the items cover a wide range of difficulty, and the median indicates that the items as a set are somewhat on the difficult side of the scale. For the 40 items, the Infit MSQs vary between 0.690 and 1.146, with a mean of 0.965 and a median 0.999. At the same time, the Outfit MSQs vary between 0.859 and 1.068 (with a mean of 0.970 and a median of 0.978). These indicate that the item fit statistics all fall within the recommended range of 0.7–1.3, and therefore, all 40 items of the scale fit the Rasch model well.

Based on the item estimates, the 40 items can be classified into three groups in terms of item difficulty. The 11 items of the 'difficult' group have facilities (*p'*s) less than 0.2. These items are separated from the rest by a natural gap in Rasch difficulties between 14.17 and 1.125. Most of these difficult items deal with some quantitative aspects of test. The remaining items deal with diagnosis, functions, assessing written expression, and above-level assessment. Generally, answering these questions correctly requires more specific training in assessment many of the teachers do not have, especially for items which are quantitative in nature.

At the other end, there are seven items in the 'easy' group, with facilities greater than 0.80 indicating that 80% or more of the teachers answered them correctly. These items are separated by a natural gap in Rasch difficulties, between −1.579 and −2.009. These items deal with concepts which can be gained through experience in assessment and are therefore commonsensical in nature; no specific training may be needed to answer such questions correctly.

In between the 'difficulty' and 'easy' groups, there are 22 items of 'appropriate' facilities, between p = 0.20 and less than p = 0.80. Their Rasch difficulties span from +1.125 to −1.579. In terms of item content, only three are from Subtest 1 (Nature and Function). There are six items from Subtest 2 (Design and Use of Test Items), seven items from Subtest 3 (Interpretation of Test Results), and six items from Subtest 4 (Reliability, Validity, and Basic Statistics). These clearly show the location of the teachers' deficits in assessment literacy.

Correlations Between Classical Facilities and Rasch Estimates

A question that has often been asked is whether the two approaches (classical and Rasch) to item analysis yield comparable results. It is therefore interesting to note that the correlation between the classical p's and the Rasch estimates is $r = |0.99|$, indicating that the two approaches of item calibration yielded almost identical results. This corroborates with many recent studies (e.g., Fan 1998; Magno 2009; Preito et al. 2003).

Discussion

Reliability

The conventional method of assessing score reliability is Cronbach's alpha coefficient, which indicates the degree of internal consistency among the items, with the assumption that the items are homogeneous. The 40 items of the scale are scored 1 (right) or 0 (wrong), and therefore, the Kuder–Richardson Formula 20 (KR20), which is a special case of Cronbach's alpha for dichotomous items, was calculated. The reliability coefficients vary from KR20 = 0.18–0.40, with a median of 0.36. Moreover, for the scale as a whole and the total sample of combined primary and secondary teachers, Cronbach's internal consistency coefficient is $\alpha = 0.471$. These indices are generally low, compared with the conventional expectation of a minimum of 0.7. This definitely leads to the question of trustworthiness.

However, there have been criticisms on Cronbach's alpha as a measure of item homogeneity or unidimensionality (Bademci 2006). One condition which might have led to the low reliabilities is the heterogeneous nature of item content among the 40 items since they cover many different aspects of educational measurement, some being qualitative and others quantitative in nature, even within a particular subtest. This renders suspect of the conventional reliability measures which assume item homogeneity. Participant homogeneity could be another factor contributing to low score reliability. Pike and Hudson (1998: 149) discussed the limitations of using Cronbach's alpha (and its equivalent KR20) to estimate reliability when using a sample with homogeneous responses in the measured construct and described the risk of drawing the wrong conclusion that a new instrument may appear to have poor reliability. They demonstrated the use of an alternate statistic that may serve as a cushion against such situation and recommended the calculation of the Relative Alpha by considering the ratio between the standard error of measurement (SEM) which itself involves the reliability as shown in the formula, thus,

$$\text{SEM} = \text{SD} * \text{SQRT}\ (1 - \text{reliability})$$

Pike and Hudson's Relative Alpha can take a value between 0.0 and 1.0 and uses an alternative way to evaluate score reliability. Their formula is,

$$\text{Relative Alpha} = 1 - \text{SEM}^2/(\text{Range}/6)^2$$

In this formula, SEM is the usual indicator of the lack of trustworthiness of the obtained scores and, under normal circumstances, the scores for a scale will theoretically span over six standard deviations. Thus, the second term on the right is an indication of the proportion of test variance that is unreliable. Relative Alpha indicates the proportion of test-variance offset for its unreliable portion, i.e., the proportion of test variance that is trustworthy.

In the present study, the maximum possible score is 40, and the theoretically possible standard deviation is 6.67 (=40/6). However, the actual data yields standard deviations of 4.24 (primary) and 4.66 (secondary) for the scale as a whole, which are 0.64 and 0.70, respectively, of the theoretical standard deviations. In other words, the two groups are found to be more homogeneous than theoretically expected.

The Relative Alphas for the primary teachers vary from 0.94 to 0.98, with a mean of 0.97. For the secondary teachers, the Relative Alphas vary from 0.93 to 0.97, with a mean of 0.97. Thus, in both cases, the statistics suggest that much of the test variance has been captured by the 40-item scale, and the scores can therefore be trusted.

Validity Evidence

Regarding content-referenced evidence, the scale was developed based on a model resulting from an analysis of empirical data and a survey of relevant literature. In addition, content analysis was conducted on the scale for a better content representation. The Rasch analysis provides further content-referenced evidence. Substantive validity evidence refers to the relationship between the construct and the data observed (Wolfe and Smith 2007a, b). In the current study, the Rasch analysis, Infit and Outfit as well as the confirmatory factor analysis provide substantive referenced evidence. Also, the alignment of the analysis based on classical test theory and the Rasch analysis supported the validity argument further.

Conclusion

This article presents preliminary evidence of the psychometric quality, the content-referenced and substantive validity of the newly developed scale. As pointed out by Popham (2006), there is a similarity between the healthcare and teaching professions in that the practitioners need to be able to properly read information about the people they serve as a prerequisite to what they intend and need to do. Thus, the importance of teachers' assessment literacy cannot be over-emphasized. There is therefore a need for an instrument that can help gauge this crucial understanding and skills of teachers. However, interest in this regard has rather a short history, and there are less than a handful of such measurement tools at our disposal at the moment.

The new scale reported here is an attempt to fill the vacuum. It covers essential conceptual skills of educational measurement which are a need-to-know for teachers if they are to perform this aspect of their profession adequately. The new scale is found to be on the 'difficult' side, partly due to a lack of relevant training among the teachers who provided the data. However, it is encouraging that its items have also been found to fit the measurement model reasonably well. What needs be done from here on is to apply the scale to larger and more representative samples of teachers in varied contexts and subjects for its consolidation. In short, the current study is the alpha, far from being the omega.

References

American Federation of Teachers, National Council on Measurement in Education, & National Education Association. (1990). *The standards for competence in the educational assessment of students.* Retrieved July 22, 2003, from http://www.unl.edu/buros/article3.html.

Bademci, V. (2006). *Cronbach's alpha is not a measure of unidimensionality or homogeneity*. Paper presented at the conference Paradigm shift: Tests are not reliable at Gazi University, 28 April 2006.

DeLuca, C., laPointe-McEwan, D., & Luhange, U. (2016). Teacher assessment literacy: A review of international standards and measures. *Educational Assessment, Evaluation and Accountability, 28*(3), 251–272.

Fan, X. (1998). Item response theory and classical test theory: An empirical comparison of their item/person statistics. *Educational and Psychological Measurement, 58*(3), 357–373.

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

Gipps, C. (1994). *Beyond testing: towards a theory of educational assessment*. London: Falmer Press.

Gotch, C. M., & French, B. F. (2014). A systematic review of assessment literacy measures. *Educational Measurement: Issues and practice, 33*(2), 14–18.

Hopkins, K. D. (1998). *Educational and psychological measurement and evaluation*. Needham Heights, MA: Allyn & Bacon.

Koh, K. H. (2011). Improving teachers' assessment literacy through professional development. *Teaching Education*, *22*(2), 255–276.

Linn, R. L., & Miller, M. D. (2005). *Measurement and assessment in teaching* (9th ed.). Upper Saddle River, New Jersey: Pearson, Merrill, Prentice Hall.

Magno, C. (2009). Demonstrating the difference between classical test theory and item response theory using derived test data. *The International Journal of Educational and Psychological Assessment, 1*(1), 1–11.

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

Pike, C. K., & Hudson, W. W. (1998). Reliability and measurement error in the presence of homogeneity. *Journal of Social Service Research, 24*(1/2), 149–163.

Popham, J. (2006). All about accountability/needed: A dose of assessment literacy. *Educational Leadership, 63*(6), 84–85. http://www.ascd.org/publications/educational-leadership/mar06/vol63/num06/Needed@-A-Dose-of-Assessment-Literacy.aspx.

Popham, W. J. (2009). Assessment literacy for teachers: Faddish or fundamental? Theory Into. *Practice, 48,* 4–11. https://doi.org/10.1080/00405840802577536.

Preito, L., Alonso, J., & Lamarca, R. (2003). Classical test theory versus Rasch analysis for quality of life questionnaire reduction. *Health and Quality of Life Outcomes, 1:27*. 10.186/1477-7525-1-27.

Rasch, G (1993). *Probabilistic models for some intelligence and attainment tests*. Chicago: Mesa Press.

Soh, K. C., & Zhang, L. (2017). The development and validation of a teacher assessment literacy scale: A trail report. *Journal of Linguistics and Language Teaching, 8*(1), 91–116.

Spolsky, B. (1978). Introduction: Linguists and language testers. In B. Spolsky (Ed.), *Approaches to language testing: Advances in language testing series* (Vol. 2, pp. V–X). Arlington, VA: Center for Applied Linguistics.

Spolsky, B. (1995). *Measured words: The development of objective language testing*. Oxford: Oxford University Press.

Stiggins, R. J. (1991). Assessment literacy. *Phi Delta Kappan, 72*(7), 534–539.

Wolf, D., Bixby, J., Glenn, J., & Gardner, H. (1991). To use their minds well: Investigating new forms of student assessment. *Review of Research in Education, 17,* 31–125.

Wolfe, E. W., & Smith, E. V., Jr. (2007a). Instrument development tools and activities for measure validation using rasch models: Part I—instrument development tools. *Journal of Applied Measurement, 8,* 97–123.

Wolfe, E. W., & Smith, E. V., Jr. (2007b). Instrument development tools and activities for measure validation using Rasch models: Part II—validation activities. *Journal of Applied Measurement, 8,* 204–233.

Author Biographies

Kay Cheng Soh (苏启祯) received Ph.D. from NUS, Singapore. He is currently Research Consultant at the Singapore Centre for Chinese Language. His academic interests include child bilingualism, creativity, world university rankings, and international achievement comparisons. His publications include books on the psychology of learning Chinese Language and various aspects of education.

Limei Zhang (张丽妹) received Ph.D. from NTU, Singapore. She is currently Lecturer at the SCCL. Her academic interests include language assessment literacy, reading and writing assessment, and learner metacognition. She has published a number of articles and book chapters on language assessment and reading and writing issues. Her most recent work is a book on the relationship between learners' metacognition and reading comprehension.