Shiken: Past and future

David Allen
allen.david[at]ocha.ac.jp

Ochanomizu University

Abstract

This article presents a history of *Shiken* since it was first published in 1997 until 2019, followed by suggestions for areas of future research in assessment to which the publication may be well suited to contribute. In the historical overview, data is presented about the following: the origins, titles, editors, and distribution; the article types; the contents of research articles and the design and methodologies they have employed. Regarding research article content, four prominent themes were identified: mass market tests, entrance exams, statistics, and validity/reliability. Regarding design and methods, research articles have tended to focus on English language tests with university students in Japan, while utilizing test and/or instrument data and quantitative methods of analysis. Recommendations for future research areas include investigations into the validity of test interpretations and uses of four-skills, vocabulary and other tests used in Japan, and language assessment literacy. Recommendations for future research design and methods include focusing more on a range of test stakeholders; various contexts, such as pre-tertiary education; and the use of qualitative and mixed methods.

Keywords: *Shiken*, TEVAL, newsletter history, journal history, four-skills tests, entrance exams

In 1996, a group of teachers and academics founded a new Special Interest Group (SIG), Testing & Evaluation (TEVAL), at the Japan Association of Language Teachers (JALT) and began a newsletter that was first published in 1997. This newsletter, *SHIKEN: JALT Testing & Evaluation SIG Newsletter*, is now in its 24th year of publication. Having reached the grand age of almost a quarter of a century, it seems fitting to present its history, to chart the ground covered so that we may look forward to new horizons of yet unexplored territory.

I was inspired to take on this task by reading a recent issue of *Assessing Writing*, in which the editor from 2002 to 2017, Liz Hamp-Lyons, and the current editor David Slomp, comment independently on the 'ideas, questions and concerns' explored in the articles featured in the journal between 2000 and the present day (Hamp-Lyons, 2019; Slomp, 2019). In the same volume, Zheng and Yu (2019) overviewed the trends in the journal and what they tell us about the field of writing assessment. These articles served as a timely reminder to me that journals are historical documents that chart the changes in academic thinking, research interests and methods. Simultaneously, as the incoming editor I needed to better understand the character of *Shiken*, particularly what has been done in the journal since its inception, in terms of the kinds of articles published and the topics covered. Thus inspired, I dug deep into the *Shiken* archives and documented my findings; the sum of which is presented in the following article.

This article is organized into two sections covering the past and the future, respectively. In the first section, I overview the contents, and highlight a number of themes that emerged from a comprehensive survey of all published issues. In terms of methodology, I adopted a straightforward approach to cataloging the digital versions of all issues and then analyzing them, using spreadsheet software. Through reading and rereading the issues, I became aware of various trends and themes, which then became subject to a more focused analysis. In this way, the texts provided the raw data, from which frequencies were tallied (e.g., the number of interviews) and categories and sub-categories were created (e.g., a category was formed for articles dealing with entrance exams, and sub-categories were formed for those that deal with this topic in a discursive manner and those that collected empirical data). I have strived to be accurate in analyzing and presenting the data, and have tried not to overindulge my own inherent biases (i.e., to my own research areas and beliefs about language teaching, learning and assessment). I have also invited and received multiple reviews from experts, some of whom have been involved in the SIG and this publication since its early days. As a result, I believe the outcome presented here is a fair and accurate representation of the publication's history.

In the second section, I look to the future of *Shiken* and in what ways the publication can contribute to language testing and evaluation in the Japan. Based on the foregoing history, and considering current trends in the field, I present a number of areas that are important for future research in the Japanese context. It is therefore hoped that this research will be of practical value in not only encouraging researchers to conduct much needed assessment research but also to submit it to *Shiken*, thereby contributing to the continued success of the publication.

The Past: *Shiken* between 1997 and 2019

General publication information

Shiken began in 1997 at the initiative of Leo Yoffe, Jeffrey Hubbell and JD Brown, and has continued up to this first issue of 2020, which will be the 24th year of publication. The publication was originally (and formally at least still is) entitled *SHIKEN*: *JALT Testing & Evaluation SIG Newsletter*. In 2012 it was temporarily renamed *Shiken Research Bulletin* but since 2014 the abbreviated title *Shiken* has been commonly used. These name changes coincided with a number of changes to the editorship, which has developed as follows: Paul Jaquith (1997 to 1998), Tim Newfields (1999 to 2011), Aaron Olaf Batty (2012), Jeffrey Stewart (2013), Trevor Holster and J. W. Lake (2014), and Trevor Holster (2015 to 2019). Although the journal was originally distributed only in printed form, all back issues were eventually made available online in HTML and PDF format by Tim Newfields. Initially, TEVAL policy was to mail printed versions to SIG members and to make back issues publicly available online one year after publication. In 2012, the TEVAL officers decided to drop printed distribution altogether and make *Shiken* an open-access, online journal, with all articles immediately available online to TEVAL members and non-members alike.

Regarding output, 49 issues in 23 volumes were published, with two issues every year for 16 years, except for two when there was only one issue (1998, 2014), and five that saw three issues per volume (2000-2003, 2009). A variety of article formats have featured with varying frequency, which is taken up in the next section.

Overview of the contents

According to the titles of the published articles, between 1997 and 2019, there were 75 *Articles*; four *Opinion Pieces*; 27 *Book Reviews*; 27 *Interviews*; 49 *Statistics Corner* articles by JD Brown; eight *Rasch Measurement in Language Education* articles; 11 *Assessment Literacy Quizzes*; one *Test Review*, though test reviews have usually been published as *Articles*; and three *Software Corner* articles. In addition, there have been a number of other sections for communicating news and events. Finally, other than during a short period during 2012 and 2015, issues have typically not featured an editorial or foreword.

From the overview of article types, a number of observations can be made. Firstly, while there have been some regular contents, particularly *Articles* and *Statistics Corner*, other article types have appeared during particular periods of time, such as *Book Reviews* and *Interviews*, which largely coincided with Newfields' editorship, and the two mini-series (i.e., the *Rasch Measurement* articles and *Assessment Literacy Quiz*). Other article types have been notably infrequent (i.e., *Opinion Pieces* and *Software Corner*). In the following, a selection of article types is described in more detail.

A special mention is required for the *Statistics Corner* articles, which have been a regular feature of *Shiken* and which for many have become synonymous with the publication. JD Brown, a founding member of the TEVAL SIG in 1997, contributed one *Statistics Corner* article every issue until his retirement in 2019. Each article responds to a question about statistics in a form that is accessible to readers with minimal experience of statistics and quantitative methods. These articles were later compiled into a book, *Statistics Corner* (2016), which is provided free-of-charge to every new member of the SIG; in other words, even after retirement, JD continues to support the next generation of language teachers and assessment researchers in Japan.

Similarly, a special note is needed for both the *Rasch Measurement in Language Education* series and the *Assessment Literacy Quizzes* contributed by Jim Sick and Tim Newfields, respectively. A series of eight articles on the topic of Rasch analysis was contributed between 2008 and 2013 by the SIG's current coordinator, Jim Sick. These articles provide an accessible introduction to Rasch measurement from an applied perspective. Similarly, Newfields' interest in developing the assessment literacy of teachers and researchers resulted in his series of quizzes between 2006 and 2011. Questions were raised on topics ranging from quantitative issues in assessment to test administration and then answered in detail with suggestions for further reading. It is noteworthy that these two series, together with the *Statistics Corner* series, indicate that a primary concern of *Shiken* has been to instruct its readers, many of whom are newcomers to the field, on methods for quantitative analysis of test-related data.

Finally, while *Research Articles* and *Book Reviews* tend to make up the bread and butter of most academic periodicals, the *Interview* series illustrates a somewhat novel aspect of *Shiken*, and also other JALT publications, such as in *The Language Teacher*. Between 2001 and 2011 interviews with language assessment specialists working in Japan and elsewhere were featured. These were mainly conducted by Newfields, and anecdotal evidence suggests that they are highly praised by the *Shiken* readership. Perhaps the reason for their popularity lies in their ability to reveal the person behind the research: only in interviews can the reader get a sense of the academic as a person who got degrees, took jobs, did research and made a career in language assessment. For readers who are ambivalent about the attraction of language assessment as a field in which to develop a career, the interviews are undoubtedly one of the most stimulating and enlightening of all the article formats. To date, 24 illustrious individuals have been interviewed, creating an incomplete who's who of contemporary language assessment research, many of them with strong ties to Japan. For the purpose of back-cataloging, but also as a reference for future interviews, here is the comprehensive, chronological list of interviewees: Leo Yoffe, Randy Thrasher, Dan Douglas, Gholamreza Hajipournezhad, Liz Hamp-Lyons, Michihiro Hirai, JD Brown (interviewed twice), Lyle Bachman, Kenji Ohtomo, Robert Gardner, Kazuhiko Saito, Yoshinori Watanabe, Michael Todd Fouts, Barry O'Sullivan, Trevor Bond, Glenn Fulcher, Carsten Roever, David Beglar, Jessica Wu, George Engelhard, Spiros Papageorgiou, Shozo Kuwata (in two languages), Alaistar van Moere (in two parts), and Meg Malone.

Articles: Content

In this section, I provide an overview of the content of *Shiken* articles. The observations presented here were made from a content analysis of all issues of the publication from 1997 to 2019. Overall, a wide range of topics are covered in the journal, though often only once or twice; these latter include vocabulary issues in assessment (Beglar, 2000; MacDonald, 2019; Trace & Janssen, 2014), test-taking strategies (Paton, Howarth & Cameron, 2018; Yoshida, 2006), ongoing assessment (Carbery, 1999; Croker, 1999), needs analysis (Kikuchi, 2005), cognitive diagnostic assessment (Aryadoust, 2011a, 2011b), rating scales (Venema, 2002) and rubric design (Duarte, 2016; Marshall, 2014). In contrast, the following four topics were addressed in multiple studies: 1) mass market tests, 2) entrance exams, 3) statistics, and 4) reliability and validity issues.

Mass market tests

A variety of mass market tests have featured in research in *Shiken* over the years, in the form of a test review, as the subject of research or as a method of assessing test-taker performance. The Test of English for International Communication (TOEIC, including TOEIC Bridge and TOEIC LPI) has been featured more than any other test, that is, eight times as the focus of the research, most recently in Paton et al. (2018), and twice as a comparison test (Hirai, 2002; Kanzaki, 2015). However, it is noteworthy that a number of these articles have been highly critical of the test itself (Chapman, 2003; Chapman & Newfields, 2008) or the organizations that administer it (McCrostie, 2010). McCrostie's (2010) article, *The TOEIC in Japan: A scandal made in heaven*, details in journalistic style the history of the Institute for International Business Communication (IIBC), which oversees test administration in Japan. In his *TOEIC®: Tried but undertested*, Chapman (2003) lamented the lack of research into the test; and Chapman and Newfields' (2008) sardonically titled *The 'New' TOEIC®* comments: 'what's remarkable about the new [2006] version of this test is how much is unaltered' (p. 33).

Only one article has focused on the development of the Test of English as a Foreign Language (TOEFL) (McNamara, 2001). However, more recently two articles have appeared on the TOEFL ITP (Institutional Testing Program), a quasi-official version of the TOEFL based on the Listening, Reading, and Structure sections of the older, paper-based TOEFL. One article focused on the interpretation of ITP scores in pretest-posttest designs (Koizumi et al., 2015) and another on teacher perceptions of the test (Collins & Miller, 2018). We predict, and hope, that given the increasingly widespread use of TOEFL ITP in Japanese higher education institutions, often for multifarious purposes (e.g., placement, progress monitoring, exit testing), research into its use, or misuse, will appear in future.

Other mass market tests have received limited attention. The International English Language Testing System (IELTS) Academic exam was featured in Aryadoust (2011b) and Boddy (2001). The Business Language Testing Service (BULATS) exam appeared once (Hirai, 2002), while the pre-two level of the EIKEN Test was reviewed by Plumb and Watanabe (2016). The Cambridge Young Learners of English Test was described by MacGregor (2001), and the American Council on the Teaching of Foreign Languages (ACTFL) Oral Proficiency Interview (OPI) was reviewed by Yoffe (1997).

Entrance exams

Entrance exams in Japan have been researched and discussed specifically in eight articles (i.e., roughly one in ten). Six of these articles provide an overview and/or discussion of the issues (Brown, 2000; Guest, 2009; Murphey, 2000/2003, 2009; Sage, 2007) and three conduct some form of empirical research into the exams (Akiyama, 2003; MacDonald, 2019; Mitchell, 2017). Regarding the discursive articles, JD Brown (2000) discussed strategies for creating positive washback from entrance exams, while in the same year an opinion piece by Murphey (2000) presented *Excerpts from an open letter submitted to the Japanese government concerning education and university entrance exams*, an article which was republished in 2003. A decade later, Murphey (2009) discussed the need to revise the system of exams for university entrance, while Guest (2009) discussed validity and reliability issues of Japanese university entrance exams. Guest's article was introduced as 'controversial' in the foreword and responses were invited, though none materialized; moreover, it was originally intended to be a two-part article, though no follow up article appeared. Finally, Sage's (2007) article critiques MEXT's 2003 Action Plan and the Ministry of Education's approach to assessing communicative competence.

Turning to the three empirical studies, Akiyama (2003) investigated the assessment of students' English speaking ability in junior high schools, the results of which are used for selection in high schools. Mitchell (2017) investigated senior high school teachers' and assistant language teachers' (ALTs') perceptions of language education pressures. Finally, MacDonald (2019) investigated the vocabulary level of the reading passages featured in the national Center Test between 2014 and 2019. In sum, it is clear that the entrance exams are a hot topic in Japan and that washback (i.e., the effect of a test on teaching and/or learning) is central to the issue. However, the majority of articles published in *Shiken* have tended to focus less on collecting primary data through empirical research and more on presenting arguments and opinions, often based on secondary sources of evidence.

Statistical and measurement issues

Eleven articles have taken up statistical issues as the primary focus of study. Six of the articles tackle these in a discussion or tutorial format, such as Smiley's (2015) account of the learning curve he encountered when studying Rasch analysis; Holster and Lake's (2015) account of using Rasch for analyzing classroom test data; and Molloy and Newfields' (2004, 2005a, 2005b) three-part article dealing with statistics in SPSS. These articles further highlight the instructional role of *Shiken*. Other studies tackle measurement issues in the format of empirical research. These include Stewart and Gibson's (2010) use of item response theory to equate pre and post-tests in the classroom; Koizumi et al.'s (2015) investigation into the regression to the mean effect as evident in TOEFL ITP scores; and Stubbe and Stewart's (2012) investigation of scoring formulas for Yes/No tests.

Reliability and validity

Perhaps the largest single, yet admittedly broad, category of articles is those that deal with the validity and reliability issues of tests. Following contemporary inclusive views on the nature of test validity (e.g., Messick, 1989; Kane, 2013; Weir, 2005), studies may be counted in this category if they either present evidence for, or discuss the nature of, any of the following aspects: construct validity (i.e., theorized dimensions of the construct being assessed), content validity (i.e., test content and format), scoring validity (i.e., reliability), criterion-related validity (i.e., concurrent and predictive validity) or consequential validity (i.e., impact and washback). At least 55 of the 79 articles appear to fall into this category, including the majority of the articles that deal with mass market tests, entrance exams, and statistical and measurement issues mentioned previously. Moreover, many articles discuss multiple aspects of validity.

It may suffice here to provide a few representative examples of validity-related studies: Jia and Zhang (2007) investigated the construct validity of an English language test for PhD applicants; Stewart, Gibson and Fryer (2012) investigated the reliability of the TOEIC Bridge test; Kanzaki (2015) investigated the validity of the Minimal English Test (see Maki, 2018) by conducting item analysis and correlational analysis with TOEIC scores; Hirai's (2002) study dealt with criterion-based validity by comparing performance scores on TOEIC and BULATS tests; and Collins and Miller (2018) investigated teachers' perceptions of TOEFL ITP, which pointed towards possible washback on teaching from the test. While these studies all dealt with a specific test, which is the most common format for articles in this category, others have sought to discuss validity-related phenomenon more generally. For instance, in opinion pieces, Roberts (2000) and Newfields (2002) both discussed the notion of face validity; Cubilo (2014) discussed the applicability of Kane's (2006) argument-based framework to classroom and program contexts; and Pan (2008) critically reviewed five washback studies.

Articles: Research design and methodologies

In this section I summarize the features of research articles in *Shiken* from the perspective of research design and methodologies. Specifically, I focus on the language being researched, macro- and micro-contexts, participants, data collection methods and methods/approaches used in analyzing data.

Every article that has focused on assessment of a particular language, has focused on English. There have been no articles about assessing other languages, such as Japanese as a second language, or French, German or Chinese as foreign languages. This is despite the fact that *Shiken*, similar to JALT and the TEVAL SIG, is not specifically focused on the English language.

In terms of the macro-context in which research was conducted, the following regions and countries have featured. Unsurprisingly, most studies were conducted in Japan (n = 47). However, there were two studies conducted in Asia: one including participants from various Asian countries (Aryadoust, 2011b) and one with a focus on innovative testing in Asia (Murphey, 2009); two studies in Australia: one discussed entrance exams in Australian universities (Gruba & Hill, 1997) and another conducted a study with participants in an Australian copper mine (Marshall, 2014); one study focused on a doctoral entrance exam in China (Jia & Zhang, 2007); one three-part study analyzed complexity ratings in Iran (Hajipournezhad, 2001, 2002a; 2002b); one collected test data in Colombia (Trace & Janssen, 2014); and two described becoming a lecturer of testing in the US (Gorsuch, 2000a, 2000b). The remainder of articles have been largely context-independent (i.e., focused on theoretical or statistical issues).

In terms of specific micro-contexts, the vast majority of studies were conducted in college/university contexts (n=42), while two were conducted in junior high schools (Akiyama, 2003; Duarte, 2016), two in high schools (Koizumi & Yano, 2019; Mitchell, 2017), and two in private companies (Hirai, 2002; Marshall, 2014). These data highlight that most research occurs at the tertiary level despite the fact that assessment is equally if not more important during pre-tertiary education. As with all academic research, this is most likely due to the relative accessibility of university participants to researchers working in tertiary institutions.

Focusing on participants, the vast majority of empirical studies recruited students, of whom the majority were at college/university (n=29), while the minority were either junior high school students (Akiyama, 2003; Duarte, 2016) or adult learners (Hirai, 2002; Marshall, 2014). One study focused on PhD applicants (Jia & Zhang, 2007), while four studies investigated teachers, two of which involved university teachers (Collins & Miller, 2018; Kikuchi, 2005) while the others involved junior high school teachers (Akiyama, 2003, as survey respondents and test interlocutors) and high school teachers and ALTs (Akiyama, 2003, as raters; Mitchell, 2017). Hajipournezhad (2002b) also reported teachers' ratings, but provided little information about the educational context.

In terms of data collection methods, of the 36 articles utilizing primary data, test performance data and/or instrument data (e.g., self-ratings on the CEFR-J in Runnels, 2013) made up the majority of data sources for analyses (n=30). Survey data featured in six articles (twice without additional data: Kikuchi, 2005; Collins & Miller, 2018; and four times with additional data sources: Akiyama, 2003; Koizumi & Yano, 2019; Mitchell, 2017; Harrison & Vanbaelen, 2013). Semi-structured interviews conducted by email were used once (Mitchell, 2017). There have been no articles in *Shiken* that have conducted and presented data from oral interviews or focus groups.

Of the 36 articles that collected primary data, 33 used quantitative analytic methods, one used qualitative analysis (Marshall, 2014), and two used mixed methods (Koizumi & Yano, 2019; Mitchell, 2017). Of all the quantitative analytic approaches, none is more synonymous with testing and assessment than item-response theory and particularly the Rasch approach. Since 1997, Rasch analysis has been used in 15 articles, beginning with Akiyama (2003) and most recently by Patterson (2019). From the data, over the 27 years of publication, the frequency of articles employing Rasch analytic techniques has increased, with all except Akiyama (2003) appearing within the last eleven years.

Only one article in *Shiken* has involved the collection and analysis of solely qualitative data: Marshall (2014) used student presentation video data to inform his development of a rating rubric for formatively assessing students' speaking performance. Additionally, two studies utilized mixed-methods. In his survey of high school teachers, ALTs' and first-year university students' perceptions of the pressures of English education at high school, Mitchell (2017) collected and analyzed quantitative survey data while following up on responses from the high school teachers in semi-structured email correspondence. In Koizumi and Yano (2019), the authors investigated an assessment of English oral presentations at a senior high school. The authors analyzed score data through quantitative (many-facet Rasch) analyses and student perceptions of the test through a survey which included Likert scale and open-ended items. Together these two studies employed different combinations of participants (i.e., high school teachers, ALTs and students), data formats (i.e., survey and email data; or test and survey data) and analyses (i.e., quantitative and qualitative, though primarily the former).

Final notes

Although space does not permit a thorough definition and analysis of the term 'quality' and how it pertains to the articles published in *Shiken*, it is certainly fair to say there has been a change in the kinds of articles published. Earlier editions had room for comic pieces, satires, and parodies, such as *Research parody: The Templin 1/2k* by Stephen Templin and Audie O'Lingual (2001). There was also the journalistic reporting of McCrostie (2010), who published a related article in the Japan Times (August 2009) entitled, *TOEIC: Where does the money go?* In contrast, more recent issues have typically only featured academic research, which has, as noted earlier, tended to be increasingly focused on quantitative analyses of test data.

The future: *Shiken* in 2020 and beyond

Based on the hitherto documented history, content and orientation of *Shiken* from 1997 to 2019, it is now possible to point to the future. In this section, I will make suggestions for the future scope and orientation of *Shiken*.

Overview

Shiken has always aimed to be a publication in which both expert and novice researchers may publish their work, and which focuses primarily, though not exclusively, on the Japan context. This aim entails a number of characteristics that distinguish *Shiken* from major journals in language testing. Particularly, *Shiken* accepts pilot studies and exploratory studies that seek to provide a basis for future research, in addition to completed research projects. Moreover, *Shiken* welcomes articles that are highly context specific and therefore not necessarily generalizable outside of Japan, or indeed even across other micro-contexts within Japan. While such articles may often be less appropriate for international journals, they constitute an important source of evidence on test use in Japan. These locally-focused articles can inform practitioners and test developers in Japan and contribute directly to the debates that are occurring in this context. However, because *Shiken* is freely available online and articles can be promoted using academic-networking sites such as researchgate.net and academi.edu, the research published therein can also attain global reach and significance.

The primary article format in *Shiken* is the peer-reviewed research article, for which some recommendations are detailed below. In addition, *Shiken* continues to consider interviews with assessment researchers and practitioners, as well as mini-series. (Readers wishing to submit interviews or a mini-series of articles are encouraged to contact the editor prior to submission).

Research areas for future studies

In this section I will highlight a number of areas for research that should be addressed in future issues of *Shiken*. These research areas are not new; in fact, they represent largely a continuation of the main themes identified previously, that is, mass market tests, entrance exams and validity studies. Perhaps most crucially, the present situation of English language education and assessment in Japan demands research into a particular combination of these subjects; that is, the use of four skills English tests for college/university admissions and the validity of the interpretations and uses of these test scores in the Japanese context. In addition, other areas include research into vocabulary tests and assessment literacy. Importantly, these suggestions are necessarily selective and are, ultimately, only suggestions. Research is required in many areas of language assessment and evaluation and *Shiken* will continue to publish a wide range of topics in accordance with the journal guidelines.

Validation of four-skills mass market tests for use as entrance exams

The main area for future research concerns the use of a selection of four skills tests for university admissions as proposed by the Ministry of Education, Culture, Sports, Science and Technology (MEXT, 2016). These include the Cambridge Assessment exams (e.g., B1 Preliminary), EIKEN exams, GTEC exams, IELTS Academic, TEAP (including TEAP CBT), and TOEFL IBT (see http://4skills.jp/index.html for up-to-date information). The purpose of this innovation is to stimulate positive washback on the learning and teaching of productive skills in pre-tertiary English education and to improve the articulation between pre-tertiary and tertiary English education. In other words, assessing all four skills equally on high-stakes entrance exams, rather than primarily reading as has been the case with the National Center Test and the various university entrance exams (i.e., *nijishiken*; Brown & Yamashita, 1996; Kikuchi, 2006), is intended to generate positive impact on English education.

Naturally, the scale of potential impact of the proposed reform has resulted in considerable discussion among both experts and non-experts alike. A noteworthy example of expert commentary is the open letter sent by the Japan Language Testing Association (JLTA, 2017) to MEXT, which praises the general aim of the innovation (i.e., positive impact) but also highlights significant shortcomings that need to be addressed if the intended impact is to be realized. For instance, much of the criticism of the proposal has focused on the significant issues of accessibility of test centers for students living in remote areas and the financial cost of taking the tests. Recently, due to harsh public criticism of the proposal, MEXT has delayed the initial implementation stage (MEXT, 2019).

The issue of using these mass market tests for Japanese university entrance purposes in the Japanese context is essentially one of validity. In other words, are the proposed interpretations and uses of test scores valid in the context of Japanese university admissions, and are the expected consequences of introducing these tests beneficial and acceptable? Although MEXT has recommended the tests, it has yet to be demonstrated that they are suitable for the purpose of Japanese university entrance. For instance, while the interpretation and use of test scores from ‘gold standard’ (Weir, 2020) international tests, such as IELTS, TOEFL and the Cambridge Assessment exams, may be valid in the context for which they were designed, such an argument does not automatically transfer to other contexts of use; in other words, the one-size-fits-all argument is not supportable (O’Sullivan, 2020; Weir, 2020). Likewise, even for locally developed tests, such as EIKEN, and for ‘glocal’ tests, such as TEAP, which have been developed locally with international expertise (Weir, 2020), evidence is required that they are fit for purpose.

To determine if a specific test is appropriate for use in a specific university admission process in Japan, evidence must be gathered that either supports or rejects the use of the test scores for that purpose in that context. Such evidence is crucial because, firstly, the recommended tests differ greatly in terms of, *inter alia*, their intended purpose, the specific skills and knowledge required, and the degree of cultural appropriateness of their content; and secondly, the contexts of use will also differ in terms of the test taker characteristics (e.g., average proficiency level) and the language needs of the university program to which they will enter. Therefore, a single broad validity argument cannot simply be made for all tests in all tertiary admissions contexts; the arguments must be test and context specific.

Although the testing agencies may be assumed to be at least partly responsible for conducting validity research, the language education and assessment community in Japan should also shoulder some of this burden. Following O'Sullivan's (2020) view on localization, local stakeholders must be involved in the process of validation if the interpretations and uses of tests are to be justified in a specific local context. Local stakeholders are most accessible to local researchers and thus there are countless opportunities for contributions by language assessment researchers in Japan. Key areas for research include:

- **Domain/Needs analyses**: It is necessary to identify the language knowledge, skills, and abilities valued in the language use situations and the relevant tasks to elicit them (Im et al., 2019). Relevant questions to ask here may include: What are the English needs in universities in Japan? How relevant and appropriate are the tasks used in the various tests to these contexts? Research has begun in this area by investigating university teachers' and students' perspectives of English language needs (e.g., Sawaki, 2016, and Tahara, 2018, respectively). However, more research is required to gain a fuller understanding of needs in various micro-contexts in Japan.
- **Curriculum alignment**: It is widely understood that assessment must be viewed as one element in a learning system, which also involves the curriculum and the delivery of that curriculum. This system should be guided by a unified approach to language learning, teaching and assessment. Considering the Japanese national educational system, the function of the to-be-replaced National Center Test was to provide an indication of achievement at the end of the national course of study. The new four-skills tests must thus also provide such an indication, which requires them to be sufficiently aligned in terms of content and level with this curriculum. Consequently, research is needed to demonstrate the extent of this alignment. The only external investigation to my knowledge is Shiratori (2018), who considered the validity of using Cambridge B1 Preliminary in admissions at Hokusei Gakuen University Junior College.
- **Uses of test scores:** Research should also investigate test developers' intended and test users' actual meanings and uses of test scores, by juxtaposing them (Im et al., 2019). In this case, questions may include: How are tests being used in various contexts in Japan? How do stakeholders interpret the scores? How are the scores being used to make decisions, and do these uses conflict with the developers' intended uses? These questions may be pursued both prior to and following implementation of the tests in specific contexts.
- **Consequences**: The primary purpose of the introduction of the four skills tests is to generate positive impact on English education in Japan. Clearly, then, research must thus be conducted into the existence, intensity, direction (i.e., positive or negative) and nature of impact. Although this research will primarily be conducted following test implementation, it can also be undertaken based on previous studies and an analysis of a context and its stakeholders in order to predict impact. As a rule, it should be assumed that both intended and unintended consequences of test use will be predicted and observed; impact research is thus essential because it may provide opportunities to intervene and promote intended positive consequences (e.g., though teacher training and assessment literacy initiatives).

Approaches to test validation

In addressing these issues, researchers will need to adopt an approach to investigating validity. The argument-based approach based on the work of Toulmin (1958) is perhaps the most widely cited approach to test validation. Variations of this approach can be seen in the work of Bachman and Palmer (2010), Chapelle, Enright and Jamieson (2008), Kane (2006, 2013), and Kunnan (2018). Bachman and Damböck (2017) provide an accessible introduction that is aimed at classroom teachers.

An alternative framework is the socio-cognitive approach, which was initially developed in the U.K. (Chalhoub-Deville & O'Sullivan, 2020; O'Sullivan, 2011, 2020; Weir, 2005) and underlies many well-established tests, including IELTS, the Cambridge Assessment exams, and TEAP, all of which are recommended four skills tests for the Japanese university admissions context. The appealing features of the socio-cognitive model include its focus on the test taker and the context, which have proved crucial in the development and validation of tests in various contexts, such as TEAP in Japan (see Dunlea et al., 2020) and the General English Proficiency Test in Taiwan (see Wu et al., 2020), as well as Aptis (see O'Sullivan, 2012) which is the first 'localizable' test in that it can be modified according to specific needs of the context.

Validation of additional test uses in Japan

In addition to the aforementioned mass-market tests for entrance purposes, some of the above research questions can be applied to other test use contexts; for example, to other tests specifically developed for university entrance, such as the newly developed British Council - Tokyo University of Foreign Studies (TUFS) Speaking Test for Japanese Universities (BCT-S; see https://www.britishcouncil.jp/exam/bct-s/about); or to tests, such as TOEFL ITP and Pearson Versant (see https://www.pearson.com/english/versant.html), which are being used for placement, progress monitoring, and/or exit testing purposes.

Another type of test that requires validation research is the vocabulary knowledge test. In recent years, a large number of vocabulary tests have been developed, including the New Vocabulary Levels Test (McLean & Kramer, 2015), described in *Shiken*. Such vocabulary tests are typically designed in accordance with one of an ever-increasing number of frequency-based wordlists. It has been argued, however, that the development and validation of many vocabulary tests has been lacking rigor and systematicity (Schmitt, Nation & Kremmel, 2019). Moreover, the approach to developing vocabulary tests that are based (exclusively) on word frequency may also be questioned. Because many teachers and researchers in Japan utilize vocabulary tests for various purposes, it is likely that this is an area to which *Shiken* can contribute.

Assessment literacy

Another key area for research is understanding and promoting the assessment literacy of teachers and other stakeholders. An interesting example of this is Berry, Sheehan and Munro (2019), who investigated the assessment literacy of teachers in Europe. Through interviews with teachers, the authors illustrated how teachers treated assessment and testing as different concepts; the former was often characterized as part of good practice in teaching, while the latter referred to formal testing, with which the teachers lacked confidence and deferred to exam boards. It would be interesting, in fact, it is essential, to investigate how teachers in Japan perceive exams (i.e., school exams, university entrance exams, and four-skills tests) in relation to the national course of study and their own teaching practices. *Shiken* has a strong history of promoting assessment literacy among its readership and thus research of this kind would be warmly received.

Research design and methodology

Turning to the nuts and bolts of assessment research, three key aspects of research design and methodology can be highlighted.

Firstly, future assessment research is needed in a greater diversity of micro-contexts, particularly at pre-tertiary levels of formal education (i.e., elementary school, junior and senior high school), but also in the private sector, particularly within the so-called shadow education system, for example, in cram schools (*juku*), prep schools (*yobikou*) and conversation schools (*eikaiwa*). Small-scale studies will likely be highly context-dependent as they are situated in specific schools or other micro-contexts, where idiosyncratic factors of participants will play a crucial role. While this is acceptable, follow-up studies in related contexts will be necessary for comparison. Larger-scale studies will require principled sampling from a number of related contexts and will likely have much greater generalizability.

Secondly, future research needs to involve a broader range of stakeholders. Regarding the most important stakeholder, the test-taker, not only university students but also test takers at various ages, especially young learners, should be the focus of future research. Moreover, although *Shiken* articles have tended, like most language assessment research, to focus on test takers and their scores, other stakeholders are also implicated in test use and therefore their perceptions and behaviors should be the subject of future research. These stakeholders include parents and guardians, employers, teachers, school principals, school administrators, school boards, examination boards, test administrators, education boards, policy makers, test developers and lawyers (Chalhoub-Deville & O'Sullivan, 2020).

Thirdly, future research should use various data collection methods and analytic approaches. The present review has shown that researchers have rarely utilized data collection methods that result in data suitable for qualitative analysis (i.e., interviews or open-ended survey items). For many areas of assessment research, such as exploring the nature of test use and consequence in society, qualitative data is fundamental. Similarly, in addition to quantitative analyses, research is needed that utilizes qualitative and mixed methods approaches.

Conclusions

This review has highlighted the scope and trends of research in *Shiken* between 1997 and 2019. It has also indicated a number of potentially fruitful avenues for future assessment research, though it should be emphasized again that these are simply suggestions rather than delimiters for research.

As one of the few specialized publications that is dedicated to assessment research in Japan, *Shiken* has played an important role in disseminating assessment research over the last almost quarter of a century. In 2020, the need for research into testing and evaluation in Japan is as imperative as it has ever been. In this context, we hope that *Shiken* will continue to serve its purpose as a vehicle for research into test validity; we hope it will continue to contribute to the important debates in language testing; and we hope through research that it will help to promote positive consequences of test use for the millions of test stakeholders in Japan.

Acknowledgements

I am grateful to the numerous reviewers who provided feedback on this paper.

References

Akiyama, T. (2003). Assessing speaking in Japanese junior high schools: Issues for the senior high school entrance examinations. *Shiken: JALT Testing and Evaluation SIG Newsletter, 7*(2), 2-11.

Aryadoust, V. (2011a). Cognitive diagnostic assessment as an alternative measurement model. *Shiken: JALT Testing and Evaluation SIG Newsletter, 15*(1), 2-6.

Aryadoust, V. (2011b). Application of the fusion model to while-listening performance tests. *Shiken: JALT Testing and Evaluation SIG Newsletter, 15*(2), 2-9.

Bachman, L. F., & Palmer, A. S. (2010). Language assessment in practice. Oxford: Oxford University Press.

Bachman, L. F., & Damböck, B. (2017). *Language assessment for classroom teachers.* Oxford: Oxford University Press.

Beglar, D. (2000). Estimating vocabulary size. *Shiken: JALT Testing and Evaluation SIG Newsletter, 4*(1), 2-5.

Berry, V., Sheehan, S., & Munro, S. (2019). What does language assessment mean to teachers? *ELT Journal*, *73*(2), 113-123.

Boddy, N. (2001). The revision of the IELTS speaking test. *Shiken: JALT Testing and Evaluation SIG Newsletter*, *5*(2), 2-5.

Brown, J. D. (2000). University entrance examinations: Strategies for creating positive washback on English language teaching in Japan. *Shiken: JALT Testing and Evaluation SIG Newsletter*, *3*(2), 2-7.

Brown, J. D. (2016). *Statistics corner: Questions and answers about language testing statistics.* Tokyo: JALT Testing and Evaluation Special Interest Group.

Brown, J. D., & Yamashita, S. O. (1995). English language entrance examinations at Japanese universities: What do we know about them? *JALT Journal, 17*(1), 7–30.

Carbery, S. (1999). Practicalities of ongoing assessment. *Shiken: JALT Testing and Evaluation SIG Newsletter*, *3*(1), 2-9.

Chalhoub-Deville, M., & O'Sullivan, B. (2020). *Validity: Theoretical Development and Integrated Arguments*. British Council Monograph Series. London & Sheffield: British Council & Equinox Publishing.

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

Chapman, M. (2003). TOEIC®: Tried but undertested. *Shiken: JALT Testing and Evaluation SIG Newsletter*, *7*(3), 2-7.

Chapman, M., & Newfields, T. (2008). The 'New' TOEIC®. *Shiken: JALT Testing and Evaluation SIG Newsletter*, *12*(2), *3*2-37.

Collins, J. B., & Miller, N. H. (2018). The TOEFL (ITP): A survey of teacher perceptions. *Shiken: JALT Testing and Evaluation SIG Newsletter 22*(2), 1-13.

Croker, R. (1999). Fundamentals of ongoing assessment. *Shiken: JALT Testing and Evaluation SIG Newsletter*, *3*(1), 10-16.

Cubilo, J. (2014). Argument-based validity in classroom and program contexts: Applications and considerations. *Shiken: JALT Testing and Evaluation SIG Newsletter*, *18*(1), 18-24.

Duarte, A. (2016). An alternative to the traditional interview test: The observed pair interview. *Shiken: JALT Testing and Evaluation SIG Newsletter 20*(2), 44-49*.*

Dunlea, J., Fouts, T., Joyce, D., & Nakamura, K. (2020). EIKEN and TEAP: How two test systems in Japan have responded to different local needs in the same context, in L.W. Su, C. Weir, & J. Wu (Eds.), *English language proficiency testing in Asia: A new paradigm bridging global and local contexts* (pp. 131-161). New York: Routledge.

Green, A. (2019). Restoring perspectives on the IELTS test. *ELT Journal, 73(2),* 207-215.

Gorsuch, G. J. (2000a). On becoming a testing teacher: Preliminary notes (Part 1). *Shiken*: *JALT Testing and Evaluation SIG Newsletter*, *3*(2), 8-17.

Gorsuch, G. J. (2000b). On becoming a testing teacher: Preliminary notes (Part 2). *Shiken*: *JALT Testing and Evaluation SIG Newsletter*, *4*(1), 9-21.

Green, A. (2007). *IELTS Washback in Context: Preparation for academic writing in higher education (Studies in Language Testing 25)*. Cambridge: Cambridge University Press.

Gruba, P., & Hill, K. (1997). An overview of Australian university entry requirements for international students. *Shiken: JALT Testing and Evaluation SIG Newsletter, 1*(2), 14-17.

Guest, M. (2008). Some new proposals and responses in ascertaining the reliability and validity of Japanese university entrance exams (part 1). *Shiken: JALT Testing and Evaluation SIG Newsletter, 12*(1), 7-13.

Hajipournezhad, G. (2001). Reading complexity judgments - Episode 1. *Shiken: JALT Testing and Evaluation SIG Newsletter, 5*(3), 2-6.

Hajipournezhad, G. (2002a). Reading complexity judgments - Episode 2. *Shiken: JALT Testing and Evaluation SIG Newsletter, 6*(1), 8-14.

Hajipournezhad, G. (2002b). Reading complexity judgments - Episode 3. *Shiken: JALT Testing and Evaluation SIG Newsletter, 6*(2), 6-9.

Hamp-Lyons, L. (2019). Reflecting on the past, embracing the future. *Assessing Writing*, 42, 100423.

Harrison, J. J., & Vanbaelen, R. (2013). Brown's approach to language curricula applied to English communication courses. *Shiken: JALT Testing and Evaluation SIG Newsletter 17*(2), 2-12.

Hirai, M. (2002). Correlations between active skill and passive skill test scores. *Shiken: JALT Testing and Evaluation SIG Newsletter, 6*(3), 2-8.

Holster, T., & Lake, J. W. (2015). From raw scores to Rasch in the classroom. *Shiken: JALT Testing and Evaluation SIG Newsletter, 19*(1), 32-41.

Im, G-H., Shin, D., & Cheng, L. (2019). Critical review of validation models and practices in language testing: their limitations and future directions for validation research. *Language Testing in Asia, 9*(14).

Jia, Y., & Zhang, W. (2007). Evaluating the construct validity of an EFL test for PhD candidates: A quantitative analysis of two versions. *Shiken: JALT Testing and Evaluation SIG Newsletter, 11*(1), 2-16.

Japan Language Testing Association (JLTA). (2017). Proposal for handling English testing within the 'Prospective University Entrance Scholastic Abilities Evaluation Test [provisional name]'. Retrieved February 14, 2020, from http://jlta2016.sakura.ne.jp/wp-content/uploads/2017/04/JLTA_proposal2017E.pdf

Kane, M. (2006) Validation. In R. Brennan (Ed.), *Educational Measurement*, 4th ed. (pp. 17-64), Westport, CT: American Council on Education and Praeger.

Kane, M. (2012). Validating score interpretations and uses. *Language Testing, 29*(1), 3–17.

Kane, M. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement, 50*(1), 1–73.

Kanzaki, M. (2015). Minimal English Test: Item analysis and comparison with TOEIC scores. *Shiken: JALT Testing and Evaluation SIG Newsletter, 19*(2), 12-23.

Kikuchi, K. (2006). Revisiting English entrance examinations at Japanese universities After a decade. *JALT Journal, 27*(1), 77–96.

Koizumi, R., In'nami, Y., Azuma, J., Asano, K., Agawa, T., & Eberl, D. (2015). Assessing L2 proficiency growth: Considering regression to the mean and the standard error of difference. *Shiken: JALT Testing and Evaluation SIG Newsletter, 19*(1), 3-15.

Koizumi, R., & Yano, K. (2019). Assessing students' English presentation skills using a textbook- based task and rubric at a Japanese senior high school. *Shiken: JALT Testing and Evaluation SIG Newsletter 23*(1), 1-33.

Kunnan, A. J. (2018). *Evaluating Language Assessments*. New York: Routledge.

MacDonald, E. (2019). An analysis of vocabulary level in reading passages of the National Center Test. *Shiken: JALT Testing and Evaluation SIG Newsletter, 32*(2), 19-27.

MacGregor, L. (2001). Testing young learners with CYLE: The new kids on the block. *Shiken: JALT Testing and Evaluation SIG Newsletter, 5*(1), *4-7.*

Maki, H. (2018). *The Minimal English Test* kenkyuu (saisho eigo tesuto): Tokyo: Kaitakusha.

Marshall, P.A. (2014). Diagnosing students' proficiency on a spoken performance assessment. *Shiken: JALT Testing and Evaluation SIG Newsletter 18*(1), 10-17.

McCrostie, J. (August, 2009). TOEIC: where does all the money go? *Japan Times*. Available at: https://www.japantimes.co.jp/community/2009/08/18/issues/toeic-where-does-the-money-go/ Last accessed 26/01/2020.

McCrostie, J. (2010). The TOEIC® in Japan: A scandal made in heaven. *Shiken: JALT Testing and Evaluation SIG Newsletter, 14*(1), 2-10.

McLean, S., & Kramer, B. (2015). The creation of a New Vocabulary Levels Test. *Shiken: JALT Testing and Evaluation SIG Newsletter, 19*(2), 1-11.

McNamara, T. (2001). The challenge of speaking: Research on the testing of speaking for the new TOEFL. *Shiken: JALT Testing and Evaluation SIG Newsletter, 5*(1), 2-3.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13–103). New York: American Council on Education & Macmillan.

MEXT (2016). *Koudai setsuzoku shisutemu kaikakukaigi: Saishu houkoku* [The final announcement of reports on discussions regarding the improvement of the upper secondary school-university articulation]. Retrieved February 14, 2020, from https://www.mext.go.jp/b_menu/shingi/chousa/shougai/033/toushin/1369233.htm

MEXT (2019). *Reiwa sannendo daigakunyugakusha sentaku ni kakaru daigaku nyushi eigo seiseki teikyo shisutemu unei taikou no haishi ni tsuite (tsuchi)* [Regarding the abolition of the operating rules for the system for providing English scores for university entrance examinations related to the selection of university enrollees in 2021]. Retrieved February 14, 2020, from https://www.mext.go.jp/a_menu/koutou/koudai/detail/1397731.htm

Mitchell, C. (2017). Language education pressures in Japanese high schools. *Shiken: JALT Testing and Evaluation SIG Newsletter, 21*(1), 1-11.

Molloy, H. P. L., & Newfields, T. (2004). Some preliminary thoughts on statistics and background information on SPSS (Part 1). *Shiken: JALT Testing and Evaluation SIG Newsletter, 8*(2), 2-5.

Molloy, H. P. L., & Newfields, T. (2005a). Some preliminary thoughts on statistics and background information on SPSS (Part 2). *Shiken: JALT Testing and Evaluation SIG Newsletter, 9*(1), 2-5.

Molloy, H. P. L., & Newfields, T. (2005b). Some preliminary thoughts on statistics and background information on SPSS (Part 3). *Shiken: JALT Testing and Evaluation SIG Newsletter, 9*(1), 2-7.

Molloy, H. P. L. (2009). Testing the test: Using Rasch person scores. *Shiken: JALT Testing and Evaluation SIG Newsletter, 13*(3), 6-12.

Murphey, T. (2000). Excerpts from an open letter to the Japanese government concerning education and university entrance exams. *Shiken: JALT Testing and Evaluation SIG Newsletter, 4*(1), 5-8.

Murphey, T. (2003). Excerpts from an open letter to the Japanese government concerning education and university entrance exams. *Shiken: JALT Testing and Evaluation SIG Newsletter, 4*(1), 5-7.

Murphey, T. (2009). Innovative school-based oral testing in Asia. *Shiken: JALT Testing and Evaluation SIG Newsletter 13*(1), 14-20.

Newfields, T. (2002). Challenging the notion of face validity. *Shiken: JALT Testing and Evaluation SIG Newsletter 6*(3), 14.

O'Sullivan, B. (2011). Language Testing, in J. Simpson (Ed.) *The Routledge Handbook of Applied Linguistics* (pp. 259-273). New York: Routledge.

O'Sullivan, B. (2012). *Aptis test development approach (ATR-1)*. Retrieved from British Council website: https://www. britishcouncil.org/sites/default/files/aptis-test-dev-approach-report.pdf

O'Sullivan, B. (2020). Localization [Foreword], in L. W. Su, C. Weir, & J. Wu (Eds.), *English language proficiency testing in Asia: A new paradigm bridging global and local contexts* (pp. xiii-xxviii). New York: Routledge.

Pan, Y-C. (2008). A critical review of five language washback studies from 1995-2007: Methodological considerations. *Shiken: JALT Testing and Evaluation SIG Newsletter, 12*(2), 2-16.

Paton, S. M., Howarth, M.W., & Cameron, A. (2018). Test-taking strategy instruction for Part 3 of the TOEIC Bridge. *Shiken: JALT Testing and Evaluation SIG Newsletter, 22*(1), 1-6.

Plumb, C., & Watanabe, D. (2016). A critique of the Grade 2 EIKEN test reading section: Analysis and suggestions. *Shiken: JALT Testing and Evaluation SIG Newsletter, 20*(1), 12-17.

Roberts, D. M. (2000). Face Validity: Is there a place for this in measurement? *Shiken: JALT Testing and Evaluation SIG Newsletter, 4*(2), 6-7.

Sage, K. (2007). MEXT's 2003 action plan: Does it encourage performance assessment? *Shiken: JALT Testing and Evaluation SIG Newsletter, 11*(2), 2-5.

Sawaki, Y. (2017). University faculty members' perspectives on English language demands in content courses and a reform of university entrance examinations in Japan: A needs analysis. *Language Testing in Asia, 7*(13).

Schmitt, N., Nation, P., & Kremmel, B. (2019). Moving the field of vocabulary assessment forward: The need for more rigorous test development and validation. *Language Teaching,* 1-12. https://doi.org/10.1017/S0261444819000326

Shiratori, K. (2019). Supporting English education reform in Japan: The role of B1 Preliminary. *Cambridge Assessment English - Research Notes: 73.* Cambridge: UCLES.

Shohamy, E. (2001). Democratic assessment as an alternative. *Language Testing, 18*(4), 373-391.

Slomp, D. H. (2019). Complexity, consequence, and frames: A quarter century of research in Assessing Writing. *Assessing Writing*, 42, 100424.

Smiley, J. (2015). Classical test theory or Rasch: A personal account from a novice user. *Shiken: JALT Testing and Evaluation SIG Newsletter, 19*(1), 16-31.

Stewart, J., & Gibson, A. (2010). Equating classroom pre- and post-tests under item response theory. *Shiken: JALT Testing and Evaluation SIG Newsletter, 14*(2), 11-18.

Stewart, J., Gibson, A., & Fryer, L. (2012). Examining the reliability of a TOEIC Bridge practice test under 1- and 3-parameter item response models. *Shiken: JALT Testing and Evaluation SIG Newsletter, 16*(2), 8-14.

Stubbe, R., & Stewart, J. (2012). Optimizing scoring formulas for yes/no vocabulary tests with linear models. *Shiken: JALT Testing and Evaluation SIG Newsletter, 16*(2), 2-7.

Tahara, T. (2018). Japanese university students' perspectives on English language needs in secondary school and university education. *Bulletin of the Graduate School of Education of Waseda University, 26*(1), 153-169.

Templin, S. A., & O'Lingual, A. (1998). Research parody: The Templin 1/2k. *Shiken: JALT Testing and Evaluation SIG Newsletter, 5*(1), 8-9.

Toulmin, S. (1958). The uses of argument. Cambridge: Cambridge University Press

Trace, J. W., & Janssen, G. (2014). Corpus-informed test development: Making it about more than word frequency. *Shiken: JALT Testing and Evaluation SIG Newsletter, 18*(1), 3-9.

Venema, J. (2002). Developing classroom specific rating scales: Clarifying teacher assessment of oral communicative competence. *Shiken: JALT Testing and Evaluation SIG Newsletter, 6*(1), 2-6.

Watanabe, Y. (1996). Does grammar translation come from the entrance examination? Preliminary findings from classroom-based research. *Language Testing, 13*(3), 318-333.

Weir, C. J. (2005). *Language test validation: An evidence-based approach*. Oxford: Palgrave.

Weir, C. J. (2020). Global, local or "glocal": Alternative pathways in English language test provision, in L. W. Su, C. Weir, & J. Wu (Eds.), *English language proficiency testing in Asia: A new paradigm bridging global and local contexts* (pp. 193-225). New York: Routledge.

Wu, R. Y-F. (2020). The General English Proficiency Test in Taiwan: Past, present and future, in L. W. Su, C. Weir, & J. Wu (Eds.), *English language proficiency testing in Asia: A new paradigm bridging global and local contexts* (pp. 9-41). New York: Routledge.

Yoffe, L. (1997). An overview of the ACTFL proficiency interview: A test of speaking ability? *Shiken: JALT Testing and Evaluation SIG Newsletter, 1*(2), 2-13.

Yoshida, K. (2006). Theoretical frameworks of testing in SLA: Processing perspectives and strategies in testing situations. *Shiken: JALT Testing and Evaluation SIG Newsletter 10*(1), 1-6.

Zheng, Y., & Yu, S. (2019). What has been assessed in writing and how? Empirical evidence from *Assessing Writing* (2000–2018). *Assessing Writing, 42,* 100421.