SHIKEN

A Journal of Language Testing and Evaluation in Japan

Volume 27 • Number 2 • December 2023

https://doi.org/10.37546/JALTSIG.TEVAL27.2

Contents

Testing and Evaluation SIG

ISSN 1881-5537

Shiken: A Journal of Language Testing and Evaluation in Japan

Volume 27 No. 2
December 2023

https://doi.org/10.37546/JALTSIG.TEVAL27.2

Language testing in changing times: An interview with Professor Daniel Isbell

Edward Schaefer[1] and Jeffrey Martin[2]
[1]edwards16@me.com
[2]jeffmjp@gmail.com
Ochanomizu University
Momoyama Gakuin University

Keywords: language testing, validity, validity argument, ethics, questionable research practices, feedback, diagnostic assessment

Daniel R. Isbell is an Assistant Professor in the Department of Second Language Studies at the University of Hawai‘i at Mānoa, where he teaches courses and supervises MA and PhD students in language assessment. He serves on the the editorial boards of the journals *Language Testing* and *TESOL Quarterly*. His primary research interest is language assessment, and he has conducted research on diagnostic assessment, self-assessment, rater effects, repeated test taking and proficiency development, test equating, specific purposes language testing, and vocabulary in context. His interests also include instructed second language acquisition, ranging from L2 pronunciation to language learning apps. It is the pleasure of the Testing and Evaluation SIG of JALT to sponsor Dr. Isbell as one of the plenary speakers at JALT PanSIG 2024.

The following interview was conducted via email correspondence to allow for a thoughtful exchange of ideas and in-depth responses to each question.

Professor Isbell, thank you for agreeing to do this interview. We'd like to start by asking you about your keynote address at PanSIG 2024. Can you give us an idea of the things you'll be discussing?

It's my pleasure to do this interview and deliver a keynote at PanSIG 2024! I have greatly enjoyed engaging with JALT's Vocabulary SIG recently, and I am very much looking forward to the PanSIG next year. I'd like to thank TEVAL for inviting me.

Language testing has really changed a lot over the last 5 to 10 years, with technology playing a much bigger role. Artificial intelligence (AI), at-home testing, automated scoring, computerized delivery of multimodal test tasks, and remote proctoring have been increasingly integrated into a wide range of language assessments, including most notably large-scale standardized tests, but we also see classroom teachers using more of these technologies, too. It is all quite exciting and also overwhelming! The conference theme of PanSIG 2024 is "Getting Back to Basics", which I plan to address by talking about how assessment basics, like validity, fairness, and practicality, can help us grapple with and appropriately evaluate major, technology-driven developments in language testing. There's a lot of hype around things like AI, but as language professionals we shouldn't lose focus of basic principles when making decisions about developing or using assessments.

One of your interests is the teaching and assessment of pronunciation, and you developed the Korean Pronunciation Diagnostic, or KPD, for your dissertation. Although it's designed as a low-stakes diagnostic assessment, you based it on a rigorous validity argument, following Kane and Chapelle. Many teachers are responsible for creating classroom assessments, but for readers who may not be familiar with the concept of validity as argument, can you say a few words about the desirability of basing even low-stakes tests on a structured validity argument?

So first, a bit about validity. Validity is something most people understand as a property of a test; people will say things like "TOEIC is not valid" or "EIKEN is valid." If you ask most people to define validity, they'd probably say things like "how accurate the test is" or "whether the test is a good measurement of something." But in language testing, and educational assessment more broadly, validity has a broader scope and includes the *use* of test scores. Our updated, more specialized definition of validity is something like *the degree to which test scores reflect the targeted knowledge, skill, or ability and are relevant and useful for making a specific decision*. So, while a test might provide a good measure of some ability, we have to think beyond that and evaluate whether that test score is appropriate for making a specific decision about a learner (e.g., what class they should be placed in, whether they have sufficient language ability to handle studying full-time in L2-medium classes).

This expanded conceptualization of validity can be hard to grapple with in practice – there's so much ground to cover. This is where validity arguments come in. Validity arguments are a framework for evaluating whether the interpretations and uses of test scores are justified. They consist of a series of inferences that we have to make when interpreting and using test scores, and each inference requires some evidence, or backing, for us to be able to accept it and advance the argument. So for example, one inference is *generalization*, which has to do with the consistency of test scores across possible conditions (e.g., different test forms, taking a test on two different days). Reliability analyses provide backing to support the inference

that *test scores are consistent*. If test scores are not consistent, it makes it hard to move the argument forward to start talking about the meaning of test scores – if you could get very different test scores on different forms of the same test, inferring something about what your test score means in terms of your English ability is premature. Validity arguments extend to test use, too, and consider inferences like *extrapolation*, the idea that test scores are predictive of performance in the real world, and *consequence implication*, the idea that using test scores to make decisions leads to desirable outcomes (e.g., student success in university, positive washback on language teaching).

Some large-scale test developers create validity arguments (Educational Testing Service, for example, but admittedly validity arguments aren't necessarily things that teachers need to create for every low-stakes test they create or use in the classroom. Kane (2013) pointed out that lower stakes demand less rigorous support. I think it's also fair to say that for most lower stakes tests, and especially those that will not be used very widely, it's okay for validity arguments to be more or less informal – most likely not even written down.

For diagnostic tools that are developed by specialists and intended to be used more broadly, however, I do think a validity argument is useful even though the stakes may be low. Validity arguments help clarify the reasoning behind test score interpretation (i.e., what a test score means about a person) and test score use (i.e., how we use a test score to make decisions and how well those decisions pan out). Validity arguments also help identify information necessary to support interpretations and uses of test scores, and in that way they double as a research agenda, helping test developers and researchers identify areas worthy of investigation. For diagnostic tools in particular, I think it's quite important for prospective users to know whether there is potential for using the scores/diagnostic feedback to support learning, which relates to inferences that assessment researchers refer to as *utilization* and *consequence implication* (or sometimes *impact*). Otherwise, there's little point in diagnosing strengths and weaknesses in the first place.

Due to the pandemic, some U.S. universities started accepting the online Duolingo English Test (DET) as an admissions test for L2 students. You did a study of the relationship between DET speaking scores and university stakeholders' evaluations of DET speaking performances. Can you tell us a little about how you applied a validity argument in order to extrapolate from DET test scores to the Target Language Use Domain (TLU), that is, success at an English medium university?

In that study (Isbell et al., 2023), we collected data relevant to what is called an *extrapolation* inference in a validity argument. This inference connects test performance to performance in the 'real world' TLU Domain. For us, we wanted to see whether DET scores and speaking performances aligned with the expectations of listeners in the real-world context of an English medium university. So, we recruited faculty, administrative staff, graduate students, and undergraduate students at our university to judge speaking task performances elicited by the DET. This approach draws heavily on the idea that 'linguistic laypersons', or non-linguists, make intuitive judgments about language performance based on their knowledge of community expectations and their own life experiences (Sato & McNamara, 2019). What we found was that test takers with higher DET scores were generally perceived by our listeners as being more comprehensible and more capable of handling increasingly demanding roles in a university – people with higher DET scores were perceived as being more capable of handling graduate studies or teaching courses.

This study does have some important limitations, though. Our approach to accessing the TLU domain was indirect, with listeners judging test-based speaking performances rather than authentic TLU performances (e.g., a real office hours interaction or class presentation). But it does lend some support to the inference that DET scores indicate differences in speaking ability that are relevant to the expectations of academic study – a preliminary but important step given that the automated scoring and simple nature of DET tasks has led to some very reasonable skepticism (see Wagner's 2020 review of the test).

Another one of your interests is ethics in language testing. In 2023, for example, you published a study reviewing developer involvement in high-stakes English proficiency tests (HSEPTs), noting that many published studies did not provide conflict of interest statements (COIs), such as ETS financing studies on TOEIC or TOEFL. You argue that the absence of COIs in published studies is an ethical lapse in the field of language testing. Can you explain what COI statements are and why they are important?

We generally expect research to provide an objective, disinterested account of some phenomenon. Conflict of interest statements are brief statements of a researcher's associations that could otherwise influence their scholarly work. Associations can be personal (e.g., a researcher's spouse or family member works for a test developer), professional (e.g., a researcher works at the university which develops and uses a placement test), or financial – it is typically the latter that comes up most frequently. Financial relationships can include regular employment (e.g., a British Council employee researching IELTS), ownership/investment (e.g., a researcher owning a part of a test developer), and other kinds of financial benefits (e.g., receiving honoraria or consulting fees). One important thing to note about COIs is that we are really talking

about *potential* COIs. We can't prove or disprove that someone's associations caused them to do anything dishonest or otherwise influenced how they conducted research and reported the results, but transparently disclosing potential COIs helps readers digest research with appropriate skepticism/criticality.

The statement for a given article is typically limited to entities addressed in the article. I've included an example disclosure statement at the end of this interview. As I have mentioned several tests and test developers, I have disclosed my relationships with relevant parties.

I believe COIs matter in language testing because test developers generally have an interest in research results that support their tests. The commercial incentives are obvious; research that strongly supports the use of an HSEPT could influence governments, universities, and other test users to accept the test, which in turn could lead to more sales. Vice versa, research findings that cast doubt on a test could have negative commercial impacts. Although research funding programs are intended to generate independently obtained evidence related to the quality or use of a test, test developers are ultimately the ones deciding which studies get funded and sometimes do (very much necessary) behind the scenes work like supply data for researchers to analyze.

In my view, it's not just commercial interests at play. Frankly, it's hard for even non-commercial test developers, such as people who develop and research placement tests for university language programs, to really be able to claim total objectivity. Test developers take pride in their work, which is a good thing, and may also want to be seen as competent professionals by their peers; research showing that their tests work well can demonstrate their skill as developers.

Some people feel that COIs in language testing research are not really needed because the relationships are obvious. But I don't feel the same way; while the connection between Duolingo and the Duolingo English Test is obvious, the connection between the Center for Applied Linguistics and the WIDA ACCESS for ELLs (a test used in U.S. public schools) is less so, especially to 'outsiders' who do not know the language testing industry very well. Peer-reviewed language testing research should strive to be something that can inform real-world test use where relevant, and to that end non-specialists should be made clearly aware of potential COIs.

In a related area, you conducted a survey of applied linguistics researchers asking them about questionable research practices (QRP). You note that QRPs exist in an ethical gray area and the label of "QRP" should not necessarily be seen as a marker of individual malintent, and you contrast this with deliberate research misconduct. You found that virtually all applied linguists who completed your survey admitted to at least one QRP (94%) and that 17% admitted to research misconduct at least once. These are disturbing findings and don't reflect well on our profession. Can you tell us the difference between QRPs and research misconduct and comment on your findings?

Research misconduct is clearly unethical, fraudulent behavior that comprises practices such as fabrication (making up data or results entirely), falsification (altering research materials, including data, with the intent to deceive), and plagiarism.

In contrast, we hope that researchers engage in research responsibly by conducting research carefully, in accordance with best practices, and transparently reporting findings. Questionable research practices are the gray area between best practices and outright misconduct. QRPs are not inherently unethical. Sometimes we engage in QRPs because there is no consensus around best practices, leading reasonable people to do things that their peers might not agree with. Sometimes QRPs happen due to carelessness or ignorance, too – it's important to keep in mind that researchers are only human. However, QRPs can also be deliberately exploited to specific ends, like obtaining a result that is statistically significant, which might cross the line into falsification in some cases (if we could somehow know the researcher's true intent, that is).

So why do misconduct and QRPs matter? Published research that is based on misconduct like fabrication of data or falsification of results is misinformation, and it can distort our understanding of language learning and teaching. This kind of fraudulent research seems fairly rare, thankfully, but it does show up every once in a while, so I think we do need to be on guard. QRPs are much more common, as our study (Isbell et al., 2022) showed. Like misconduct, QRPs can potentially distort our understanding. One common QRP is excluding non-significant findings from a study. Let's say you conduct a study on a hot topic like ChatGPT in L2 writing instruction, and you want to examine grammatical accuracy, lexical diversity, and fluency (length). Compared to a control condition, you might find that students who are allowed to use ChatGPT produce texts with significantly fewer grammatical errors. So you report those findings, which are quite exciting, and maybe you can get it published. But let's say you also examined fluency, in terms of text length, and lexical diversity, and found that there weren't any significant differences between the ChatGPT and the control condition. If you don't report those findings, too, you end up with an unbalanced picture.

In the context of research in Diagnostic Language Assessment (DLA), you highlighted the importance of actionable feedback, emphasizing its adaptability in both quantitative and qualitative formats. Considering the diverse spectrum of L2 learners in terms of proficiency, identifiable strengths and weaknesses, age, language learning goals, and motivation,

individualized feedback becomes crucial. Could you provide insights into which types of learners stand to benefit the most from diagnostic feedback, given these varying factors?

This is a difficult question, and one that, to my knowledge, there's not a clear, universal answer from empirical research. So, my answer should be taken with a grain of salt.

One thing that seems to be important is the ability to understand and 'digest' the feedback. Feedback that is rather granular and detailed can be quite helpful, but it might be inaccessible if it is delivered in the target language rather than a learner's L1 (or other proficient language). Also important is the motivation of the learner – not so much in capital-M motivation associated with big-picture theories, but specific motivation to really engage with feedback provided by a diagnostic procedure and do the work to address any weak areas. It is often the case that learners will look at some feedback (whether it is diagnostic results or any other kind of feedback) but only pay attention to overall results and quickly move on, as they might lack a specific drive or desire to really work on specific language features.

Harding et al., (2015) and Isbell (2021) propose that for a test to be diagnostic, its results need to be in the form of feedback that is relevant and actionable in subsequent L2 learning. Concerning the question of which language skills or competencies are able to be diagnosed, first, suppose a framework for language use was applied, such as Bachman's model of Communicative Competence, what skills or competencies do you think are currently best diagnosable? What are examples of such tests that currently exist? Second, to what extent could diagnostic tests be developed to assess and provide feedback for the other components that are not yet well tested, according to language models?

I am very much sympathetic to Harding et al.'s (2015, and Alderson et al., 2015) argument that subcomponents of communicative language skills, particularly those that can be assessed in a discrete, granular manner, are most readily diagnosable. Their view clearly influenced my design for the KPD, which drills down into the perception and production of individual phonemes. While L2 pronunciation (and speaking ability more broadly) depends on more than just segmental aspects of pronunciation, phonemes do matter in every spoken utterance. Ideally, there would be other diagnostic tools to dig deeper into specific weaknesses of suprasegmental pronunciation features, but I had to start somewhere.

Clark and Endres (2021) is a nice example of a diagnostic assessment of English grammar targeted at the A2 proficiency level. Grammar, by itself, is not really so important as a communicative language learning outcome – except that it is something we draw on *constantly* when using language to communicate. Hence the inclusion of Grammatical Competence in Bachman's model. So Clark and Endres' grammar diagnostic, which is all presented in the written modality, is something that might help teachers and learners understand in greater detail why they have difficulty understanding when reading or have difficulty expressing some ideas clearly when writing.

I do think there's room to expand on diagnostic assessment practices through more formalized screening or observation procedures. Alderson and colleagues do discuss this and allude to some ways that less formal observations can motivate more detailed diagnostic tools being used, but in my own research and experience (including supervising student projects related to DLA), the first step of figuring out who might benefit from additional diagnostic procedures is really key. So in this area, coarser grained diagnostic screening/observation of more communicative, even integrated skills seems necessary and, I think, is quite possible.

Thank you for taking the time to answer these questions. We're looking forward to hearing your keynote address at the JALT PanSIG in May 2024.

Declaration of competing interests:
D. Isbell has received research funding from British Council, Duolingo, Educational Testing Service, and Pearson, honoraria from Educational Testing Service, and has consulted for IELTS UK and Duolingo.

References

Alderson, J. C., Brunfaut, T., & Harding, L. (2015). Towards a Theory of Diagnosis in Second and Foreign Language Assessment: Insights from Professional Practice Across Diverse Fields. *Applied Linguistics*, *36*(2), 236–260. https://doi.org/10.1093/applin/amt046

Bachman, L. F. (1990). *Fundamental Considerations in Language Testing*. Oxford University Press.

Clark, T., & Endres, H. (2021). Computer-based diagnostic assessment of high school students' grammar skills with automated feedback – an international trial. *Assessment in Education: Principles, Policy & Practice*, *28*(5–6), 602–632. https://doi.org/10.1080/0969594X.2021.1970513

Harding, L., Alderson, J. C., & Brunfaut, T. (2015). Diagnostic assessment of reading and listening in a second or foreign language: Elaborating on diagnostic principles. *Language Testing*, *32*(3), 317–336. https://doi.org/10.1177/0265532214564505

Isbell, D. R. (2021). Can the Test Support Student Learning? Validating the Use of a Second Language Pronunciation Diagnostic. *Language Assessment Quarterly*, 1–26. https://doi.org/10.1080/15434303.2021.1874382

Isbell, D. R., Crowther, D., & Nishizawa, H. (2023). Speaking performances, stakeholder perceptions, and test scores: Extrapolating from the Duolingo English test to the university. *Language Testing*. Online advance publication. https://doi.org/10.1177/02655322231165984

Isbell, D. R., Brown, D., Chen, M., Derrick, D. J., Ghanem, R., Arvizu, M. N. G., Schnur, E., Zhang, M., & Plonsky, L. (2022). Misconduct and questionable research practices: The ethics of quantitative data handling and reporting in applied linguistics. *The Modern Language Journal*, 24. https://doi.org/10.1111/modl.12760

Isbell, D. R., & Kim, J. (2023). Developer involvement and COI disclosure in high-stakes English proficiency test validation research: A systematic review. *Research Methods in Applied Linguistics*. https://doi.org/10.1016/j.rmal.2023.100060

Kane, M. T. (2013). Validating the Interpretations and Uses of Test Scores: Validating the Interpretations and Uses of Test Scores. *Journal of Educational Measurement*, *50*(1), 1–73. https://doi.org/10.1111/jedm.12000

Sato, T., & McNamara, T. (2019). What Counts in Second Language Oral Communication Ability? The Perspective of Linguistic Laypersons. *Applied Linguistics*, *40*(6), 894–916. https://doi.org/10.1093/applin/amy032

Wagner, E. (2020). Duolingo English Test, Revised Version July 2019. *Language Assessment Quarterly*, 1–16. https://doi.org/10.1080/15434303.2020.1771343

Conducting a Rasch Analysis in jMetrik

Trevor A. Holster
trevor@fukujo.ac.jp
Fukuoka Jogakuin University

Abstract

This step-by-step guide to conducting a Rasch analysis using the jMetrik software package describes how to format and import data, conduct a simple classical test theory item analysis, and then conduct a simple Rasch analysis. Some basic graphical outputs are also described. After completing the steps described in this guide, novice users should be able to conduct a basic analysis unaided and be able to explore the more advanced features of jMetrik by referring to the manual.

Keywords: jMetrik, Rasch analysis, item analysis

Batty's (2023) introductory guide included a tutorial on installing jMetrik (Meyer, 2018) and creating a scoring key. This follow-up tutorial builds on the basics explained there so readers should work through that introduction first. The appendices contain a brief tutorial on alternative methods of creating and editing an answer key that novices may find simpler than the advanced method described by Batty (2023). More advanced analyses and detailed technical explanations are provided in the jMetrik manual (Meyer, 2014), readers who intend to use jMetrik on a regular basis are recommended to purchase that manual.

Before you begin, you will need to download the practice dataset (Holster, 2023). You can then create a new database and import the data into jMetrik (Appendix A.) This data came from a pilot administration of a proposed placement test. You will then need to create an answer key, either following Batty's (2023) guide, or the alternative methods described in Appendices B to D. A sample copy of the test and a brief explanation of each section is provided in Appendix E.

Generating Student Scores

For classroom use and for most research purposes, you will need to generate students' scores from your test. Once you have imported the data and created an answer key, click **Transform >>> Test Scaling.** In the *Score* section, select *Sum Score* as the score type. Give this a name, for example "total". Select the items you want to include in the test score. You can select all items using CTRL + A or you can select a range of items by holding down the SHIFT key when you click on the first and last item. Click the *Run* button. In the Data tab, click *Refresh Data View* at the top. There is a new variable on the right named "total" that shows the total number correct for each student. To calculate the percentage scores, select *Average Score* in the *Score* section of the *Test Scaling* dialogue box. I'll call this "percentage". Select all the items and click "Run". Refresh the data display.

Exporting and saving results

Click **Manage >>> Export Data.** Make sure that "Comma" is selected from the Delimiter panel. Select CSV Files as the file type and give the file a meaningful name, "Test Scores", for example. This will save a new data file with the students' scores included. These scores can then be used for assigning student grades or further analyzed for research purposes.

Test Reliability

Journal editors and reviewers will require you to report descriptive statistics and a test reliability coefficient. Click **Analyze >>> Item Analysis**. Select all the items and select *Polyserial correlation* in the Item-total Correlation type. Click *Run* and the results will appear in a new tab. You can save this as a text file from the File menu by clicking "Save" or "Save As". *Test Level Statistics* are provided at the bottom of the output, reproduced in Table 1. This includes descriptive statistics and a *Reliability Analysis* section, reproduced in Table 2. The coefficient alpha of .91 is excellent for a classroom test. As a rule-of-thumb, values below .80 would generally be considered low, while high-stakes decisions would typically require values of .90 or higher.

Table 1

Test level statistics

Number of Items = 130
Number of Examinees = 281
Min = 38.0000
Max = 114.0000
Mean = 78.9929
Median = 80.0000
Standard Deviation = 13.9817
Interquartile Range = 18.0000
Skewness = -0.0427
Kurtosis = -0.2297
KR21 = 0.8480

Table 2

Reliability analysis

Method	Estimate	95% Conf. Int.	SEM
Guttman's L2	0.9120	(0.8967, 0.9260)	4.1558
Coefficient Alpha	0.9062	(0.8900, 0.9212)	4.2888
Feldt-Gilmer	0.9093	(0.8937, 0.9238)	4.2172
Feldt-Brennan	0.9091	(0.8933, 0.9236)	4.2233
Raju's Beta	0.9062	(0.8900, 0.9212)	4.2888

Classical Item Analysis

The *Item Analysis* output can be used to identify poorly functioning items. Table 3 shows statistics for four items, LC1, LC2, CE12, and V42. The *Discrimin* column shows the item discrimination. We would usually want items with discriminations of .40 or higher. The *Difficulty* column in Table 3 shows the proportion of correct responses, in this case, 0.8968. In this classical analysis, a higher value for difficulty indicates an easier item because more students answered correctly. Item LC1 has a low discrimination of .15, meaning that it does a poor job of discriminating between high and low-proficiency students. In the case of LC1, the item is very easy. Only 10% of students did not answer correctly so the reason for the low discrimination is probably that the item is too easy for this sample of students. However, it does have a positive discrimination and is the first item in the test so retaining it would be justified on the grounds that it provides a confidence boost to begin the test.

Item LC2 has a reasonably good discrimination of .54 and difficulty of 0.72, so this item is functioning acceptably, but is fairly easy for these students. Item CE12 has a slightly negative discrimination of -.04, and is also very easy. Looking at the sample test form, it is a cloze elide shadowing item, but we can see that it is an *unplanned item*. The unplanned CE items are necessary for the functioning of this test format because they function as distractors, but they are very poor at discriminating between high and low-ability students.

Item V42, shown below, also has a discrimination very close to zero and is a very difficult item, with only 14% of students answering correctly. It is a 5-option multiple-choice question so we would expect 20% of students to answer correctly through random guessing. This indicates a malfunctioning distractor. The response option E, a distractor, attracted 49% of responses. Even more concerning, the distractor E had a discrimination of .15 compared to .03 for the item key, option B. The problem with this item is related to the distractor E, *Highlight*. It is possible that this item would function well with a sample of students of higher proficiency, but this distractor should be replaced in future administrations of this test.

Sample Item (Item V42)

Look

A) Echo B) Peer C) Rent D) Exclude E) Highlight

Table 3

Sample Item Statistics

Item	Option	Score	Difficulty	Std. Dev.	Discrimin.
lc1	Overall		0.8968	0.3048	0.1517
	C	0.0	0.0036	0.0597	-0.4529
	E	0.0	0.0036	0.0597	0.0480
	G	0.0	0.0036	0.0597	-0.0952
	I	0.0	0.0107	0.1030	-0.4952
	N	0.0	0.0036	0.0597	-0.5483
	O	1.0	0.8968	0.3048	0.1517
	Q	0.0	0.0676	0.2515	-0.0440
	X	0.0	0.0107	0.1030	-0.3511
lc2	Overall		0.7224	0.4486	0.5442
	A	0.0	0.0107	0.1030	-0.4051
	B	0.0	0.0071	0.0842	-0.1026
	C	0.0	0.0391	0.1943	-0.3644
	D	0.0	0.0178	0.1324	-0.3994
	E	0.0	0.0142	0.1187	-0.0275
	F	0.0	0.0107	0.1030	-0.3331
	G	1.0	0.7224	0.4486	0.5442
	H	0.0	0.0427	0.2026	-0.3895
	I	0.0	0.0071	0.0842	-0.5912
	J	0.0	0.0071	0.0842	-0.3470
	K	0.0	0.0071	0.0842	-0.1155
	Q	0.0	0.0285	0.1666	-0.0965
	R	0.0	0.0107	0.1030	-0.3871
	S	0.0	0.0071	0.0842	-0.3598
	T	0.0	0.0036	0.0597	-0.0713
	X	0.0	0.0641	0.2453	-0.4746
ce12	Overall		0.8968	0.3048	-0.0428
	0.0	1.0	0.8968	0.3048	-0.0428
	1.0	0.0	0.1032	0.3048	-0.0309
v42	Overall		0.1388	0.3463	0.0250
	A	0.0	0.1708	0.3770	-0.0055
	B	1.0	0.1388	0.3463	0.0250
	C	0.0	0.0605	0.2388	-0.5323
	D	0.0	0.1246	0.3308	-0.2168
	E	0.0	0.4947	0.5009	0.1537
	X	0.0	0.0107	0.1030	-0.1979

Rasch Analysis

Rasch analysis converts percentage scores into log-odds units, or *logits,* which provide equal interval measures. This is desirable for researchers and also allows person ability and item difficulty to be mapped onto the same measurement

scale. Novices to Rasch analysis should refer to Sick's series of introductory articles (Sick, 2008a, 2008b, 2009a, 2009b, 2010, 2011, 2013a, 2013b) or to *Applying the Rasch Model* (Bond & Fox, 2015) a standard introductory text.

Conducting Rasch Analysis

Click **Analyze >>> Rasch Models**. From the Global tab, select all the items. The *Center on items* box is checked by default and the *Linear Transformation* section has defaults of Mean = 0, Scale = 1, and Precision = 4. Leave all of these on the default setting. In the Item tab, check *Save item estimates* and enter a meaningful name for the output table. I will call mine *Rasch Items.* In the Person tab, check *Save person fit statistics* and *Save person estimates*. Then click *Run*.

Rasch Item Analysis

Rasch item analysis focuses on *fit statistics* rather than just item correlations, and item difficulty is given in logits. By default, mean item difficulty is set to 0.00 logits, with a useful range of item difficulty usually from -3 logits (a very easy item) to 3 logits (a very difficult item.) Mean-square fit statistics have a mean value of approximately 1.00, with values greater than 1.50 indicating a level of misfit (or underfit) requiring investigation. Fit statistics are reported as a weighted mean-square (WMS), or *infit,* statistic that reflects when items are well matched to person ability, and an unweighted mean-square (UMS), or *outfit*, statistic that reflects outlying responses.

Table 4 shows the fit statistics for the four items we looked at in the CTT analysis. Item LC1 is extremely easy, with a logit value of -1.68. The WMS value of 1.05 shows good infit, but the UMS value of 1.54 shows a level of outfit that is of concern. This is probably because the item is very easy and some high-ability students answered incorrectly, perhaps just a single student. Item LC2 is moderately easy and slightly over-fitting, with mean-square fit values below 1.00. This is consistent with what we saw in the classical item analysis.

Table 4

Final JMLE item statistics

Item	Difficulty	Std. Error	WMS	Std. WMS	UMS	Std. UMS
lc1	-1.68	0.20	1.05	0.38	1.54	2.23
lc2	-0.37	0.14	0.91	-1.35	0.83	-1.83
ce12	-1.68	0.20	1.10	0.71	1.95	3.53
v42	2.75	0.18	1.15	1.27	1.45	2.33

Item CE12 is also very easy and very badly misfitting, with a UMS value of 1.95. This is an unplanned cloze elide item and nearly all students succeeded on this this item. The very high UMS value will be because some high-ability students gave incorrect responses. Item V42 is extremely difficult, with a logit value of 2.75 and is also moderately misfitting, with a UMS value of 1.45. This item had a badly functioning distractor and very few students answered correctly. The misfit is probably the result of high-ability students being confused by the bad distractor, but low-ability students succeeding through lucky guessing.

Discrimination and data-model fit in Rasch analysis

Discrimination in Rasch analysis does not mean item correlations as in classical analysis. Discrimination means the slope of the *item characteristic curve* (ICC). The Rasch model is based on the assumption that all items have equal discrimination. Items with higher discrimination will *overfit* the model, items with lower discrimination will *underfit*, or misfit the model. Items with negative correlations will usually badly misfit, but higher item correlations do not automatically mean better fit to the Rasch model. Negative correlations will always indicate a problem, but low positive correlations are not a problem as long as the fit statistics are acceptable.

The mean-square fit statistics have an expected value of 1.00, which indicates well-fitting responses. Values larger than 1.50 indicate a level of misfit that is of concern and values above 2.00 indicate serious problems. Values below 1.00 indicate *overfit*, which means that the item is more consistent than expected. The *weighted mean-square* (WMS) statistic, or *infit*, is weighted to exaggerate responses where the person ability was well-matched to the item difficulty. These

responses provide more information than responses where items are much higher or lower than person ability. A high infit value usually indicates a serious problem with the item.

The *unweighted mean-square* (UMS)statistic, or *outfit*, is not information weighted. High outfit values reflect outlying responses, where high-ability students fail on easy items or low-ability students succeed on difficult items. A common cause of high outfit is that items are extremely easy or extremely difficult, so a very small number of responses can cause extreme items to misfit.

jMetrik also provides standardized fit statistics, which show whether the results are statistically significant, with values outside the range of -2.0 to 2.0 indicating statistical significance ($p < .05$). Statistical significance is largely a result of the sample size, so if we have a lot of students, nearly all item misfit will be statistically significant, even if it is not substantively large. The mean-square statistic is therefore much more useful because it shows the substantive size of the misfit, but it is not the only consideration.

Rasch Person Diagnosis

On the data screen in jMetrik, click *Refresh Data View.* You will now see extra columns of data giving the Rasch analysis results for each student. Click **Manage >>> Export Data** to export this data as a file that can be open in Excel.

Person ability scores

Person ability is called *theta* in Rasch analysis and is reported in logits, not percentage scores. This makes it easy to compare student ability and item difficulty because they are both given in logits. In the data window, we can see a variable called *sum,* another one called *vsum*, and then *theta.* The *sum* score is the total number correct. We can see that the *vsum* score is always 3 less than the *sum* score. This is because there are three items that every student succeeded on in this test. This is shown in the Rasch output for items CE6, CE40, and V1, where these items are flagged as *Minimum.* Items (or persons) with extreme scores (i.e. 0% or 100%) do not provide any information for the analysis so they are not used in the estimation of Rasch logits.

Rasch person fit statistics

Rasch analysis provides WMS (infit) and UMS (outfit) statistics for persons as well as for items. We can use these for student diagnostics. Table 5 shows the most misfitting students. Four of these students have very high scores and one has a very low score of 38. Student S4426 has a very high UMS (outfit) value of 3.41. This is a very high-ability student, with a total of 114 and a logit ability of 2.98. This high-ability student has failed on some easy items so we would look at their test to see why. For example, they may have been confused by the cloze elide section and made careless errors on that section.

Table 5

Rasch Person Statistics

St_No	Total	Sum	Vsum	Theta	Stderr	Extreme	Wms	Stdwms	Ums	Stdums
S4426	114	114	111	2.98	0.31	No	1.07	0.43	3.41	2.04
S2304	111	111	108	2.72	0.29	No	0.89	-0.64	2.60	1.74
S4409	108	108	105	2.48	0.28	No	1.14	0.95	2.43	1.77
S4436	107	107	104	2.40	0.27	No	1.14	1.00	2.27	1.68
S2110	78	78	75	0.64	0.23	No	1.05	0.44	2.04	3.03
S2403	38	38	35	-1.58	0.25	No	1.35	2.51	2.27	2.53

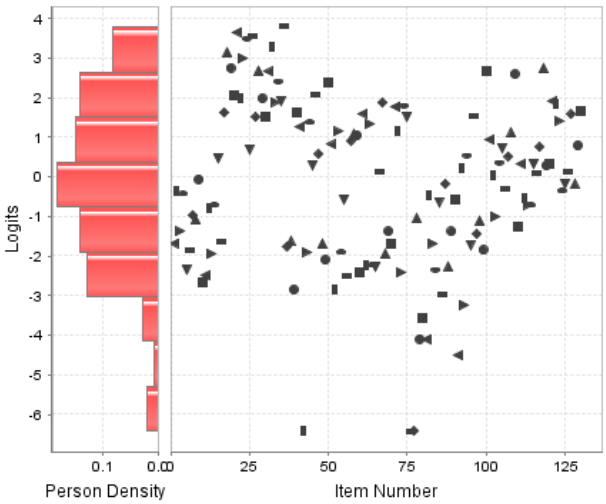
Student S2403 is a very low-ability student, with a total of 38 and logit ability of -1.58. This student has a very high outfit mean-square (UMS) value of 2.27. This student has succeeded on some difficult items so we would look at their test to try to understand why. For example, they may have struggled with the listening sections but performed much better on the vocabulary section. We would need to look at their test to confirm this, but that is the type of pattern we would look for in diagnosing misfitting students.

Creating a Person-Item Map

Go to the *Raschitems* data screen. Click **Graph >>> Item Map.** Select *bparam* and in the *Item Parameter Table* box select *Rasch Items.* Click **Run.** jMetrik will produce a person-item map. This compares student ability to item difficulty on the same vertical logit scale, as shown in Figure 1. Mean item difficulty is set to 0.00 logits so a student with ability of 0.00 logits has a 50% chance of succeeding on an item of mean difficulty. The range of item difficulty is quite well matched to the range of student difficulty overall. However, there are very few items below -3 logits so the test cannot measure the very lowest ability students. The item map can be saved by right-clicking and selecting *Save as* to save the image as a .png file suitable for journal publication.

Figure 1

Person-item map. The vertical scale shows the logit scale of item difficulty and person ability.



Comparing the test sections

Cloze listening: This section was relatively easy, so it is suitable for separating very low ability students who need remedial instruction from mainstream students. Revised test specifications are needed to address this.

Cloze dictation: The cloze dictation section was much more difficult, so this section is suitable for identifying students who would benefit from more challenging extension classes.

Cloze elide shadowing: This section has a large gap between the relatively difficult planned items and the relatively easy unplanned items. This format is confusing for low-ability students, so it is only suitable for higher ability students and Japanese language instructions are needed, along with a practice test to familiarize students with the format.

Vocabulary synonymy: These items span a very large range of difficulty, with many easy items near the start of the section and more difficult items towards the end. It would be desirable to replace some medium and high difficulty items with extremely easy items to target remedial students.

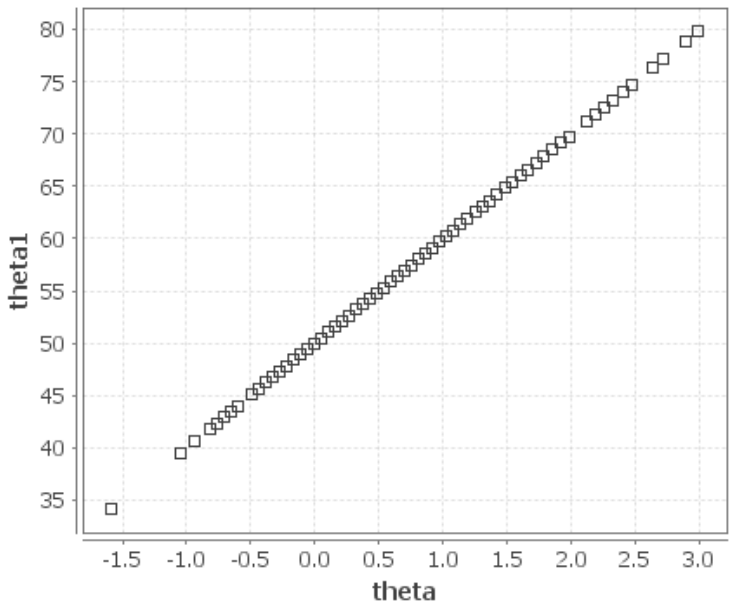
Rescaling Rasch Logit Scores

Logit measures of person ability are confusing to most people. We can rescale them to more convenient units. Click **Analyze >>> Rasch Models** and select all the items for analysis. In the *Linear Transformation* section, set *Mean* to 50 and *Scale* to 10. In the *Person* tab, check the box for *Save person estimates.* This will rescale the logit scores to have a mean item difficulty of 50, with 1 logit rescaled to 10 scaled units. In the data screen, click *Refresh Data View.* There will now be a new theta column that shows student ability on a scale that resembles percentage scores.

Click **Graph >>> Scatterplot.** Select theta (the original logit scores) for the X axis and theta1 (the rescaled scores) for the Y axis, then click *Run.* The scatterplot, shown in Figure 2, shows a perfectly linear transformation of the scores.

Figure 2

Rescaled logit scores. The logit scores have been rescaled to a more user-friendly scale with mean item difficulty of 50 and 1 logit equal to 10 scaled units.



Rasch Reliability

In the output from the Rasch analysis, there is a table called *Scale Quality Statistics,* reproduced in Table 6. This includes a reliability coefficient for both items and persons, plus separation and strata indices. The person reliability of .91 is analogous to the Cronbach alpha statistic. The separation index is calculated from the reliability coefficient. The figure of 3.16 means that we can be confident that there are three statistically distinct levels of person ability. What this means is that we have very high confidence that the highest students are actually more proficient than the average students, and also that the average students are actually more proficient than the lowest ability students. In the case of this test, we could confidently use it as a placement test to separate students into two or three different course levels.

Table 6

Scale quality statistics

Statistic	Items	Persons
Observed Variance	3.5695	0.6242
Observed Std. Dev.	1.8893	0.7901
Mean Square Error	0.0433	0.0568
Root MSE	0.2080	0.2383
Adjusted Variance	3.5263	0.5674
Adjusted Std. Dev.	1.8778	0.7533
Separation Index	9.0278	3.1607
Number of Strata	12.3705	4.5476
Reliability	0.9879	0.9090

Item reliability

Rasch analysis also provides an item reliability statistic, plus item separation and strata indices. In this case, the item reliability is .98, with a separation index of 9.03. We have very high confidence that the most difficult items are actually

more difficult than the easiest items. This test was intended to include some very easy items and some much more difficult items, so the item reliability suggests that this was achieved. Item reliability is of limited use for classroom teachers, but for research projects where measurement of task difficulty of tasks is required, the item reliability is the more important consideration.

Conclusion

This guide is intended to help novices conduct a simple Rasch analysis. Rasch analysis provides diagnostic tools that are unavailable with classical analysis of percentage scores, but the concepts underlying Rasch analysis are quite different from classical analysis. Logit measures of ability/difficulty are confusing to teachers and students who are accustomed to percentage scores, but they are invaluable for test developers and researchers because person ability and item difficulty can be mapped onto the same scale. A further source of confusion is that classical analysis of item discrimination is based on item correlations, with higher correlations assumed to indicate a better functioning item, while Rasch analysis focuses on fit statistics, with a completely different definition of discrimination.

References

Batty, A. O. (2023). jMetrik Guide. *Shiken*, *27*(1), 15-29.

Bond, T., & Fox, C. M. (2015). *Applying the Rasch model: Fundamental measurement in the human sciences* (3rd ed.). Routledge.

Davies, A. (1967). The English proficiency of overseas students. *British Journal of Educational Psychology*, *37*(2), 165-174. https://doi.org/10.1111/j.2044-8279.1967.tb01925.x

Holster, T. A. (2017). Cloze-elide as a classroom reading test. *Shiken*, *21*(2), 1-19. http://teval.jalt.org/sites/teval.jalt.org/files/21_02_01_Holster_Cloze_Elide.pdf

Holster, T. A. (2023). *jMetrik practice data*. https://hosted.jalt.org/teval/sites/jalt.org.teval/files/jMetrik%20Data.xls

Meyer, J. P. (2014). *Applied Measurement with JMetrik*. Routledge. https://books.google.co.jp/books?id=Xd-8mQEACAAJ

Meyer, J. P. (2018). *jMetrik 4.11*. In https://itemanalysis.com/jmetrik-download/

Sick, J. (2008a). Rasch measurement in language education: Part 1. *Shiken*, *12*(1), 1-6. https://hosted.jalt.org/test/PDF/Sick1.pdf

Sick, J. (2008b). Rasch measurement in language education: Part 2: Measurement scales and invariance. *Shiken*, *12*(2), 26-31. https://hosted.jalt.org/test/PDF/Sick2.pdf

Sick, J. (2009a). Rasch measurement in language education: Part 3: The family of Rasch Models. *Shiken*, *13*(1), 4-10. https://hosted.jalt.org/test/PDF/Sick3.pdf

Sick, J. (2009b). Rasch measurement in language education: Part 4: Rasch analysis software programs. *Shiken*, *13*(3), 13-16. https://hosted.jalt.org/test/PDF/Sick4.pdf

Sick, J. (2010). Rasch Measurement in language education: Part 5: Assumptions and requirements of Rasch measurement. *Shiken*, *14*(2), 23-29. https://hosted.jalt.org/test/PDF/Sick5.pdf

Sick, J. (2011). Rasch measurement in language education: Part 6: Rasch measurement and factor analysis. *Shiken*, *15*(1), 15-17. https://hosted.jalt.org/test/PDF/Sick6.pdf

Sick, J. (2013a). Rasch measurement in language education: Part 7: Judging plans and disjoint subsets. *Shiken*, *17*(1), 27-32. https://hosted.jalt.org/sites/jalt.org.teval/files/SRB-17-1-Sick-RMLE7_0.pdf

Sick, J. (2013b). Rasch measurement in language education: Part 8: Rasch measurement and inter-rater reliability. *Shiken*, *17*(2), 23-26. https://hosted.jalt.org/sites/jalt.org.teval/files/SRB-17-2-Sick-RMLE8.pdf

Appendix A: Creating a Database and Importing Data

Open jMetrik. Click **Manage** >>> **New Database**. Name your database using lowercase letters. I will name my new database "jaltdemo." Open the database by clicking **Manage** >>> **Open Database**.

Download the sample data file (Holster, 2023). Unfortunately, the JALT server will not host comma separated values (.csv) files so the data is provided in an Excel file that needs to be converted to .csv format using the *Save As* option in Excel. Click **Manage** >>> **Import Data**, then browse to the sample database and select the jMetrik Data.csv file as the data file. Give it a table name. I will call mine jaltdemo.

Click **Import** to import the data. This data is now saved in the new database and will be there the next time you open this database in jMetrik.

Data Layout

The first column is student codes, running from S1201 to S5334. There are 281 students.

The top row is item codes. The test has four sections, with 130 items:

1. Items 1-16: Coded LC (listening cloze) 16 items Response codes A to T
2. Items 17-36: Coded LD (listening dictation) 20 items Response codes A to T
3. Items 37-76: Coded CE (cloze elide) 40 items Response codes 0 and 1
4. Items 77-130: Coded V (vocabulary) 54 items Response codes A to E

Missing data

Missing data should be coded NA. The sample dataset does not contain missing data.

Appendix B: Creating an Answer Key Using Basic Item Scoring

Click **Transform >> Basic Item Scoring**. The top row shows the item codes. The answer key for each item needs to be entered into the second row and the number of response options needs to be entered into the third row. We just need to copy the row of answer keys from Table B1 and B2 into the second row of the wizard and enter 24 in the third row for every item.

These two listening sections use response codes of A to T, plus X for skipped items. In this case, we want to score X as an incorrect response, so the number of response options should be 24, to include all the letters from A to X. The Tab key will let you move to the next box in the row. You can use CTL * C to copy the contents of a cell and CTL +V to paste it into another cell. Once all the data entry for sections 1 and 2 is complete, click OK to save the answer key.

Table B1

Listening cloze answer key

Item	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
Code	LC	LC	LC	LC	LC	LC	LC	LC	LC	LC	LC	LC	LC	LC	LC	LC
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
Key	O	G	B	C	J	L	P	S	T	M	K	F	R	I	N	D

Table B2

Listening dictation answer key

Item	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36
Code	LD	LD	LD	LD	LD	LD	LD	LD	LD	LD	LD	LD	LD	LD	LD	LD	LD	LD	LD	LD
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
Key	I	H	L	J	Q	C	D	N	O	P	B	R	M	G	S	F	K	E	T	A

Appendix C: Creating an Answer Key Using Advanced Item Scoring

Items with the same answer key can be processed together using the Advanced Item Scoring wizard. Table C1 lists all the vocabulary items with the answer key A, so we can enter all these together. This is much faster than entering each one separately.

Table C1

Vocabulary answer key A

Item	84	91	93	94	100	109	114	115	116	119	122	123	124	125	127
Code	V8	V15	V17	V18	V24	V33	V38	V39	V40	V43	V46	V47	V48	V49	V51
Key	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A

Click **Transform >> Advanced Item Scoring.** First, enter all the characters from A to E, plus X in the "Option" column, In the "Score" column, enter a score of 1 for the option A, and a score of 0 for all the other options. Select item V8 and click the > button to move it to the selection panel. On a Windows computer, you can hold down the CTRL key to select multiple items, so you can select all the items listed above and move them to the selection panel. On a Mac, you use the Command button for this. Finally, click "Submit'. You will see the selected items highlighted in bold and the scoring syntax at the bottom. Then click OK to save the answer key for these items. Now we need to repeat this for all the items with "B" as the scoring key (Table C2), then "C" (Table C3), "D" (Table C4), and finally "E" (Table C5), with the correct response scored as 1 and the other responses scored as 0.

Table C2

Vocabulary answer key B

Item	82	89	96	104	105	107	118	130
Code	V6	V13	V20	V28	V29	V31	V42	V54
Key	B	B	B	B	B	B	B	B

Table C3

Vocabulary answer key C

Item	78	79	81	83	86	87	92	110	120	126	128
Code	V2	V3	V5	V7	V10	V11	V16	V34	V44	V50	V52
Key	C	C	C	C	C	C	C	C	C	C	C

Table C4

Vocabulary answer key D

Item	77	80	85	88	95	99	101	102	103	106	111	112	121	129
Code	V1	V4	V9	V12	V19	V23	V25	V26	V27	V30	V35	V36	V45	V53
Key	D	D	D	D	D	D	D	D	D	D	D	D	D	D

Table C5

Vocabulary answer key E

Item	90	97	98	108	113	117
Code	V14	V21	V22	V32	V37	V41
Key	E	E	E	E	E	E

Appendix D: Scoring Numerically Coded Rating Scale Items

Section 3 of the test uses a format called *cloze elide* (CE). These use a numeric key, not an alphabetic one. Some items (*unplanned items*) are reverse scored, they are listed in Table D1. Open the Advanced Item Scoring wizard and enter "0" and "1" as the response options, with scores of 1 and 0 respectively (i.e. response option "0" has a score of 1; response option "1" has a score of 0). The *planned items* use a more regular scoring key, with a response of "1" scored as 1 and a response of "0" scored as zero. These are listed in Table D2. Use the Advanced Scoring wizard to score these, with a response of "0" having a score of 0 and a response of "1" having a score of 1.

Table D1

Cloze elide unplanned items answer key

Item	37	38	39	42	43	48	49	52	54	55	56	60	62	64	65	68	69	70	73	76
Code	CE 1	CE 2	CE 3	CE 6	CE 7	CE 12	CE 13	CE 16	CE 18	CE 19	CE 20	CE 24	CE 26	CE 28	CE 29	CE 32	CE 33	CE 34	CE 37	CE 40
Key	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Table D2

Cloze elide planned items answer key

Item	40	41	44	45	46	47	50	51	53	57	58	59	61	63	66	67	71	72	74	75
Code	CE 4	CE 5	CE 8	CE 9	CE 10	CE 11	CE 14	CE 15	CE 17	CE 21	CE 22	CE 23	CE 25	CE 27	CE 30	CE 31	CE 35	CE 36	CE 38	CE 39
Key	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1

Scoring Polytomous Items

The sample data only contains dichotomous items, with scores of 0 or 1. Items with polytomous scales can be scored using the advanced item scoring wizard in much the same way as these dichotomous items.

Appendix E: An Overview of this Test

This test was developed to demonstrate that classroom tasks could be adapted to make a very cheap placement test that could be printed on a single A3 sheet and administered in less than 60 minutes. The test has four sections, with 130 items.

Listening Part 1: Listening cloze

This section, coded "LC" in the data file, forms a testlet, with the 16 items sharing 20 response options and response codes from A to T. The key (i.e. the correct response) for each item functions as a distractor for the other items, potentially introducing dependency between items. There are four additional distractors to reduce this effect. Students listened to the recorded conversation two times and were instructed to write the missing words in the gaps. This is a cloze format test because it is possible to identify the correct responses (i.e. "cloze" the gap) just by reading. It was intended as a listening test for very low-proficiency students who cannot complete cloze format reading items.

Listening Part 2: Partial Dictation

This section also forms a testlet, coded "LD", with the 20 items sharing 20 response options and response codes from A to T. The key (i.e. the correct response) for each item functions as a distractor for the other items, potentially introducing dependency between items. Students listened to the recorded conversation two times and were instructed to write the missing words in the gaps. This is a partial dictation format test because it is difficult to identify the correct responses just by reading. It was intended as a listening test for higher proficiency students who need more difficult items than the cloze listening format.

Listening Part 3: Cloze Elide Shadowing

Part 3 is development of a rarely used format called cloze elide (Davies, 1967; Holster, 2017). Items are coded "CE" in the data file, with dichotomous responses of 0 and 1. Some words have been added at random places in the text. Students were required to read and listen to the text (i.e. shadow the listening) and cross out (i.e. elide) the extra words. This was intended as a format that would sharply discriminate between very low-proficiency students and average students. In this example, some lines of text have one extra word (called planned items) and other lines of text do not have any extra words (called unplanned items). Students must shade the answer bubble for planned items and leave it unmarked for unplanned items. This makes the test machine scoreable, but limits the number of items compared with the more common format of treating each word as a separate item and allowing multiple planned items per line of text (Davies, 1967).

Part 4: Vocabulary Synonymy

Part 4 of the test uses a vocabulary synonymy format, coded "V" in the dataset, with response codes from A to E. The item stem consists of a single word, with five response options. One response option (the item key) is a synonym of the stem. The other four responses (the distractors) are not synonymous.

学籍番号

Family name: _____

First name: _____

Listening Part 1

A: ANA skydiving.
B: Yeah, hi, ah… this is the skydiving (1)_____?
A: Yes sir, what can we (2)_____ for you?
B: Well, it's my brother's (3)_____ next week and we, ah… my sister and me, we want to (4)_____ him skydiving lessons.
A: Skydiving lessons or a (5)_____?
B: What do you (6)_____?
A: Well, we have lessons and we also have (7)_____ jumps.
B: Jumps?
A: Yes, tandem jumps, you jump (8)_____ with an instructor.
B: Together with an instructor?
A: Yes, and (9)_____ also have lessons.
B: I see… um… how (10)_____ is the single jump, um… with the instructor?
A: That's a hundred dollars sir.
B: A hundred dollars. Ok, uh… how much are the (11)_____?
A: Well, it's a two day (12)_____, and it costs three hundred dollars.
B: Three hundred dollars, I see. Ok, well I (13)_____ we'll go with the single jump with the (14)_____. I'll check with my sister and (15)_____ you back.
A: Ok, thanks for (16)_____.

A) bathroom	B) birthday	C) buy	D) calling	E) cars
F) course	G) do	H) fun	I) instructor	J) jump
K) lessons	L) mean	M) much	N) phone	O) school
P) single	Q) thanks	R) think	S) together	T) we

1.	A	**B**	C	D	E	F	G	H	I	J	K	**L**	M	N	**O**	P	O	R	S	T
2	A	B	C	D	E	F	**G**	H	I	J	**K**	L	M	N	O	P	Q	R	S	T
3	A	**B**	C	D	E	F	G	H	I	J	**K**	L	M	N	O	P	Q	R	S	T
4	A	B	**C**	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T
5	A	B	C	D	E	F	G	H	I	**J**	K	L	M	N	O	P	Q	R	S	T
6	A	B	C	D	E	F	G	H	I	J	K	**L**	M	N	O	P	Q	R	S	T
7	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	**P**	Q	R	S	T
8	A	B	**C**	D	E	F	G	H	I	J	K	**L**	M	N	O	P	Q	R	**S**	T
9	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	**T**
10	A	B	C	D	E	F	G	H	I	J	K	L	**M**	N	O	P	Q	R	S	T
11	A	B	C	D	E	F	G	H	I	J	**K**	L	M	N	O	P	Q	R	S	T
12	A	B	C	D	E	**F**	G	H	I	J	K	L	M	N	O	P	Q	R	S	T
13	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	**R**	S	T
14	A	B	C	D	E	F	G	H	**I**	J	K	L	M	N	O	P	Q	R	S	T
15	A	B	C	D	E	F	G	H	I	J	**K**	L	M	**N**	O	P	Q	R	S	T
16	A	B	C	**D**	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T

Listening Part 2

When translating xxxx (1)_____ xxxx language into another, the xxxx (2)_____ xxxx is always present. One often sees rather funny translations of Chinese dishes or Japanese xxxx (3)_____ xxxx around the world. However, sometimes translation errors xxxx (4)_____ xxxx and stay around long enough so xxxx (5)_____ xxxx becomes accepted. An example of this is the English name for the popular Mexican dish, refried beans. Refried xxxx (6)_____ xxxx common Mexican dish served in the United States and Canada. They are often xxxx (7)_____ xxxx dip for chips or as a side dish for a main meal. However, refried beans xxxx (8)_____ xxxx refried. The English name refried came about due to a translation error. This xxxx (9)_____ xxxx be called well-fried beans, not refried beans. In Spanish, these refried beans are called *frijoles refritos*. In Spanish *frijoles* means *beans* and *fritos* is an xxxx (10)_____ xxxx *fried*. So, *frijoles fritos* would mean *fried beans*. The translation error xxxx (11)_____ xxxx of the prefix *re-* in the word *refritos*. In English, the prefix *re-* usually means that you do something again, for example, to retake a test xxxx (12)_____ xxxx it again, to redesign a pamphlet is to design it again, to remake a xxxx (13)_____ xxxx make a movie again. However, in Spanish, the prefix *re-* does not mean xxxx (14)_____ xxxx again, instead it is used to emphasize something. So, in this case, xxxx (15)_____ xxxx to emphasize that the beans xxxx (16)_____ xxxx fried, not fried again. However, this mistaken xxxx (17)_____ xxxx and *frijoles refritos* continue xxxx (18)_____ xxxx *refried beans*. Perhaps a bit of good advice would be to always xxxx (19)_____ xxxx knowledgeable native speaker before you xxxx (20)_____ xxxx.

A) a	B) about	C) are	D) as	E) be
F) been	G) do	H) for	I) from	J) get
K) has	L) in	M) is	N) not	O) should
P) that	Q) the	R) to	S) was	T) with

1.	Ⓐ	Ⓑ	Ⓒ	Ⓓ	Ⓔ	Ⓕ	Ⓖ	Ⓗ	Ⓘ	Ⓙ	Ⓚ	Ⓛ	Ⓜ	Ⓝ	Ⓞ	Ⓟ	Ⓠ	Ⓡ	Ⓢ	Ⓣ
2	Ⓐ	Ⓑ	Ⓒ	Ⓓ	Ⓔ	Ⓕ	Ⓖ	Ⓗ	Ⓘ	Ⓙ	Ⓚ	Ⓛ	Ⓜ	Ⓝ	Ⓞ	Ⓟ	Ⓠ	Ⓡ	Ⓢ	Ⓣ
3	Ⓐ	Ⓑ	Ⓒ	Ⓓ	Ⓔ	Ⓕ	Ⓖ	Ⓗ	Ⓘ	Ⓙ	Ⓚ	Ⓛ	Ⓜ	Ⓝ	Ⓞ	Ⓟ	Ⓠ	Ⓡ	Ⓢ	Ⓣ
4	Ⓐ	Ⓑ	Ⓒ	Ⓓ	Ⓔ	Ⓕ	Ⓖ	Ⓗ	Ⓘ	Ⓙ	Ⓚ	Ⓛ	Ⓜ	Ⓝ	Ⓞ	Ⓟ	Ⓠ	Ⓡ	Ⓢ	Ⓣ
5	Ⓐ	Ⓑ	Ⓒ	Ⓓ	Ⓔ	Ⓕ	Ⓖ	Ⓗ	Ⓘ	Ⓙ	Ⓚ	Ⓛ	Ⓜ	Ⓝ	Ⓞ	Ⓟ	Ⓠ	Ⓡ	Ⓢ	Ⓣ
6	Ⓐ	Ⓑ	Ⓒ	Ⓓ	Ⓔ	Ⓕ	Ⓖ	Ⓗ	Ⓘ	Ⓙ	Ⓚ	Ⓛ	Ⓜ	Ⓝ	Ⓞ	Ⓟ	Ⓠ	Ⓡ	Ⓢ	Ⓣ
7	Ⓐ	Ⓑ	Ⓒ	Ⓓ	Ⓔ	Ⓕ	Ⓖ	Ⓗ	Ⓘ	Ⓙ	Ⓚ	Ⓛ	Ⓜ	Ⓝ	Ⓞ	Ⓟ	Ⓠ	Ⓡ	Ⓢ	Ⓣ
8	Ⓐ	Ⓑ	Ⓒ	Ⓓ	Ⓔ	Ⓕ	Ⓖ	Ⓗ	Ⓘ	Ⓙ	Ⓚ	Ⓛ	Ⓜ	Ⓝ	Ⓞ	Ⓟ	Ⓠ	Ⓡ	Ⓢ	Ⓣ
9	Ⓐ	Ⓑ	Ⓒ	Ⓓ	Ⓔ	Ⓕ	Ⓖ	Ⓗ	Ⓘ	Ⓙ	Ⓚ	Ⓛ	Ⓜ	Ⓝ	Ⓞ	Ⓟ	Ⓠ	Ⓡ	Ⓢ	Ⓣ
10	Ⓐ	Ⓑ	Ⓒ	Ⓓ	Ⓔ	Ⓕ	Ⓖ	Ⓗ	Ⓘ	Ⓙ	Ⓚ	Ⓛ	Ⓜ	Ⓝ	Ⓞ	Ⓟ	Ⓠ	Ⓡ	Ⓢ	Ⓣ
11	Ⓐ	Ⓑ	Ⓒ	Ⓓ	Ⓔ	Ⓕ	Ⓖ	Ⓗ	Ⓘ	Ⓙ	Ⓚ	Ⓛ	Ⓜ	Ⓝ	Ⓞ	Ⓟ	Ⓠ	Ⓡ	Ⓢ	Ⓣ
12	Ⓐ	Ⓑ	Ⓒ	Ⓓ	Ⓔ	Ⓕ	Ⓖ	Ⓗ	Ⓘ	Ⓙ	Ⓚ	Ⓛ	Ⓜ	Ⓝ	Ⓞ	Ⓟ	Ⓠ	Ⓡ	Ⓢ	Ⓣ
13	Ⓐ	Ⓑ	Ⓒ	Ⓓ	Ⓔ	Ⓕ	Ⓖ	Ⓗ	Ⓘ	Ⓙ	Ⓚ	Ⓛ	Ⓜ	Ⓝ	Ⓞ	Ⓟ	Ⓠ	Ⓡ	Ⓢ	Ⓣ
14	Ⓐ	Ⓑ	Ⓒ	Ⓓ	Ⓔ	Ⓕ	Ⓖ	Ⓗ	Ⓘ	Ⓙ	Ⓚ	Ⓛ	Ⓜ	Ⓝ	Ⓞ	Ⓟ	Ⓠ	Ⓡ	Ⓢ	Ⓣ
15	Ⓐ	Ⓑ	Ⓒ	Ⓓ	Ⓔ	Ⓕ	Ⓖ	Ⓗ	Ⓘ	Ⓙ	Ⓚ	Ⓛ	Ⓜ	Ⓝ	Ⓞ	Ⓟ	Ⓠ	Ⓡ	Ⓢ	Ⓣ
16	Ⓐ	Ⓑ	Ⓒ	Ⓓ	Ⓔ	Ⓕ	Ⓖ	Ⓗ	Ⓘ	Ⓙ	Ⓚ	Ⓛ	Ⓜ	Ⓝ	Ⓞ	Ⓟ	Ⓠ	Ⓡ	Ⓢ	Ⓣ
17	Ⓐ	Ⓑ	Ⓒ	Ⓓ	Ⓔ	Ⓕ	Ⓖ	Ⓗ	Ⓘ	Ⓙ	Ⓚ	Ⓛ	Ⓜ	Ⓝ	Ⓞ	Ⓟ	Ⓠ	Ⓡ	Ⓢ	Ⓣ
18	Ⓐ	Ⓑ	Ⓒ	Ⓓ	Ⓔ	Ⓕ	Ⓖ	Ⓗ	Ⓘ	Ⓙ	Ⓚ	Ⓛ	Ⓜ	Ⓝ	Ⓞ	Ⓟ	Ⓠ	Ⓡ	Ⓢ	Ⓣ
19	Ⓐ	Ⓑ	Ⓒ	Ⓓ	Ⓔ	Ⓕ	Ⓖ	Ⓗ	Ⓘ	Ⓙ	Ⓚ	Ⓛ	Ⓜ	Ⓝ	Ⓞ	Ⓟ	Ⓠ	Ⓡ	Ⓢ	Ⓣ
20	Ⓐ	Ⓑ	Ⓒ	Ⓓ	Ⓔ	Ⓕ	Ⓖ	Ⓗ	Ⓘ	Ⓙ	Ⓚ	Ⓛ	Ⓜ	Ⓝ	Ⓞ	Ⓟ	Ⓠ	Ⓡ	Ⓢ	Ⓣ

Listening Part 3

Inventions

An invention is a new technology that makes our lives easier. Some important inventions were: the compass, steel, nails, the printing press, the lens, steam engines, electricity, and the telegraph.

The compass uses a magnetic arrow theoretical that points north and south. It allows explorers to find their way across long distances. The first compasses were invented sense in China about 200 BC.

Steel is made from iron and carbon. It was invented in Turkey more than 4 000 years ago and was used to make hammer knives. Steel was very expensive until inventors in England developed cheap steel in the 17th century. This made England very powerful trap.

Nails are identification metal pins that are used to join wood together. The first nails were hand-made and were expensive inspiration. In the 19th century, inventers learned how to make cheap nails using machines. The first machine made nails were cut from sheets of steel, but later they were made from wire.

The printing press is a machine for making many catch copies of books and magazines. Before the printing press, books were written shark by hand. This was very slow and expensive. Woodblock printing was invented in China about 200 AD. This uses a block of wood which is cut away to make all the letters and pictures fighter for one page. Movable type was invented in China in the 11th century. Each character was cut into a small block of wood. The characters were then put together to make a page of text. Johannes Gutenberg invented the modern printing press in 1439. This used metal blocks for each letter and oil-based ink. Books and newspapers became much cheaper after jeans this.

A lens is a manufacturing clear material like glass that bends light. They were used in ancient Greece to make fire. Lenses can make things look bigger charter or smaller. They are used to make reading glasses, telescopes, microscopes, and many other scientific instruments tolerance. Galileo Galilei used lenses to make a telescope in 1609. He used his telescope to look at the moon and planets.

Steam engines use energy from boiling chunk water to power a machine. The first steam powered pumps were used in the 17th century to pump water out of coal mines. Steam engines were very powerful, so factories became bigger and cheaper. In the 19th century, steam engines were used to power the first trains, so transport became much faster thread.

Electricity is a form shared of energy that occurs naturally as lightning. Electrical generators were invented in the 19th century. After that, electricity was used to power factories, lights, elevators, and many other machines. Modern cities would be impossible without electricity.

The telegraph was used to indigenous send messages through wires before telephones were invented. The first telegraphs were difficult conscience to use, but Samuel Morse invented a much simpler telegraph in 1844. In 1850 an undersea cable was used to broadcast connect London to Paris. Before this, communication over long distances was very slow, but the telegraph made it possible to communicate with other countries instinct in a few minutes.

1. Good
(A) Big
(B) Old
(C) High
(D) Great
(E) Small

2. Take
(A) Sit
(B) Live
(C) Bring
(D) Write
(E) Happen

3. Small
(A) Long
(B) Black
(C) Little
(D) Important
(E) Political

4. Study
(A) Set
(B) Lead
(C) Stop
(D) Learn
(E) Follow

5. Talk
(A) Stop
(B) Allow
(C) Speak
(D) Create
(E) Follow

6. Let
(A) Read
(B) Allow
(C) Speak
(D) Spend
(E) Create

7. Minute
(A) Art
(B) Girl
(C) Moment
(D) Result
(E) Research

8. Get
(A) Buy
(B) Die
(C) Love
(D) Wait
(E) Consider

9. Come
(A) Pull
(B) Decide
(C) Report
(D) Return
(E) Explain

10. Call
(A) Oil
(B) Data
(C) Phone
(D) Piece
(E) Practice

11. Sure
(A) Easy
(B) Recent
(C) Certain
(D) General
(E) Personal

12. Girl
(A) Fire
(B) South
(C) Future
(D) Daughter
(E) Population

13. Go in
(A) Note
(B) Enter
(C) Share
(D) Shoot
(E) Reduce

14. Lose
(A) Hang
(B) Laugh
(C) Prove
(D) Design
(E) Forget

15. Black
(A) Dark
(B) Heavy
(C) Middle
(D) Specific
(E) Beautiful

16. Difficult
(A) Dark
(B) Civil
(C) Tough
(D) Beautiful
(E) Commercial

17. Day
(A) Date
(B) Beach
(C) Spirit
(D) Element
(E) Feature

18. President
(A) Chief
(B) Crowd
(C) Prison
(D) Survey
(E) Target

19. More
(A) Fair
(B) Warm
(C) Light
(D) Additional
(E) Responsible

20. People
(A) Ears
(B) Folk
(C) Bands
(D) Affairs
(E) Lessons

21. After
(A) Chief
(B) Direct
(C) Native
(D) Dangerous
(E) Following

22. Sport
(A) Tour
(B) Danger
(C) Finding
(D) Plastic
(E) Exercise

23. Bad
(A) Busy
(B) Smart
(C) Initial
(D) Terrible
(E) Surprised

24. Want
(A) Lack
(B) Dance
(C) Match
(D) Accuse
(E) Succeed

25. Certainly
(A) Barely
(B) Closer
(C) Typically
(D) Definitely
(E) Unfortunately

26. Special
(A) Proud
(B) Liberal
(C) Massive
(D) Unusual
(E) Increased

27. Study
(A) Drag
(B) Wrap
(C) Knock
(D) Review
(E) Display

28. Scream
(A) Wash
(B) Shout
(C) Bother
(D) Recover
(E) Purchase

29. Vote
(A) Urge
(B) Elect
(C) Quote
(D) Benefit
(E) Deserve

30. Temperature
(A) Hit
(B) Knife
(C) Missile
(D) Climate
(E) Participation

31. Steel
(A) Dust
(B) Iron
(C) Tale
(D) Advance
(E) Independence

32. Purpose
(A) Pace
(B) Peer
(C) Butter
(D) Approval
(E) Intention

33. Pot
(A) Pan
(B) Lord
(C) Viewer
(D) Teaspoon
(E) Administrator

34. Game
(A) Prize
(B) Virus
(C) Match
(D) String
(E) Stranger

35. Amazing
(A) Rapid
(B) Solar
(C) Efficient
(D) Incredible
(E) Sophisticated

36. Screen
(A) Cow
(B) Mama
(C) Infant
(D) Monitor
(E) Congressman

37. Test
(A) Bull
(B) Heel
(C) Complex
(D) Horizon
(E) Examination

38. Write
(A) Print
(B) Sense
(C) Wander
(D) Devote
(E) Appoint

39. Terrible
(A) Awful
(B) Organic
(C) Distinct
(D) Subsequent
(E) Agricultural

40. Nice
(A) Lovely
(B) Weekly
(C) Homeless
(D) Continued
(E) Institutional

41. Okay
(A) Sacred
(B) Related
(C) Artistic
(D) Concrete
(E) Acceptable

42. Look
(A) Echo
(B) Peer
(C) Rent
(D) Exclude
(E) Highlight

43. Healthy
(A) Fit
(B) Unlike
(C) Minimum
(D) Written
(E) Occasional

44. Fix
(A) Pump
(B) Value
(C) Repair
(D) Minimize
(E) Volunteer

45. Go
(A) Age
(B) Hook
(C) Peel
(D) Vanish
(E) Stumble

46. Shelf
(A) Rack
(B) Ounce
(C) Costume
(D) Sympathy
(E) Declaration

47. Angry
(A) Upset
(B) Immune
(C) Anonymous
(D) Associated
(E) Experienced

48. Similar
(A) Alike
(B) Someday
(C) Upstairs
(D) Partially
(E) Consequently

49. Famous
(A) Known
(B) Teenage
(C) Short-term
(D) Cooperative
(E) Preliminary

50. Water
(A) Bee
(B) Dam
(C) Liquid
(D) Mushroom
(E) Catalogue

51. Rack
(A) Tray
(B) Jungle
(C) Patent
(D) Videotape
(E) Speciality

52. Forest
(A) Tray
(B) Aisle
(C) Jungle
(D) Patent
(E) Specialty

53. Get
(A) Cure
(B) Boast
(C) Color
(D) Inherit
(E) Distract

54. Stupid
(A) Atop
(B) Dumb
(C) Neat
(D) Valid
(E) Intact

Call for Papers

Shiken: A Journal of Language Testing and Evaluation in Japan is seeking submissions for future issues on an ongoing basis. After checking the guidelines for publication below, please send your submission to the editor at tevalpublications@gmail.com.

Overview

Shiken aims to publish articles concerning language assessment issues relevant to assessment professionals, researchers, classroom practitioners and language program administrators. *Shiken* is dedicated to publishing articles on assessment in various educational contexts in and around Japan.

Article formats include research papers, replication studies, and review articles, all of which should typically not exceed 7000 words; as well as informed opinion pieces, technical advice articles, and interviews, all of which should typically not exceed 3000 words. Please contact the editor if you wish to submit an article that differs from these formats.

Novice researchers are encouraged to submit, but should aim for papers that focus on a single main issue. Please review recent issues of *Shiken* to understand the level of detail, depth of discussion and provision of empirical evidence that is required for acceptance.

Format

To facilitate double-blind peer review, the name(s) of the author(s) should not be mentioned in the manuscript. All citations and references to the author or co-author's work should read (Author, Year) e.g. "(Author, 2017)."

Articles should be formatted according to the guidelines of the *Publication Manual of the American Psychological Association (APA), 7th Edition*. Submissions should be formatted in Microsoft Word (.docx format) using 12-point Times New Roman font. The page size should be set to A4, with a 2.5 cm margin. The body text should be block justified, and single-spaced. Paragraphs should be separated by a blank line space. Each new section of the paper should have a section heading. Please review the most recent issues of *Shiken* for examples of section headings.

Research articles must be preceded by an abstract that succinctly summarizes the article content in less than 200 words. The reference section should begin on a new page immediately following the body text. Authors are responsible for comprehensive and accurate referencing including adding DOI or URL information wherever possible. Tables and figures should be numbered and titled. Separate sections for tables and figures should follow the references. Any appendices should be numbered and should follow the tables and figures.

Evaluation

All papers are double-blind peer-reviewed by two expert reviewers. Initial evaluation is usually completed within four weeks. The whole process from submission to publication is normally completed within six months.