SHIKEN

A Journal of Language Testing and Evaluation in Japan

Volume 25 • Number 1 • June 2021

https://doi.org/10.37546/JALTSIG.TEVAL25.1

Contents

Testing and Evaluation SIG
ISSN 1881-5537

Shiken: A Journal of Language Testing and Evaluation in Japan

Volume 25 No. 1
June 2021

https://doi.org/10.37546/JALTSIG.TEVAL25.1

Voices in the field: An interview with Nick Saville

By David Allen
Ochanomizu University
https://doi.org/10.37546/JALTSIG.TEVAL25.1-1

Bio

Nick Saville is currently the Director of Thought Leadership at Cambridge Assessment. He studied Linguistics at the University of Reading and holds a PhD from CRELLA (Center for Research in English Language Learning and Assessment) at the University of Bedfordshire in language test impact. He is currently Secretary General of the Association of Language Testers in Europe (ALTE). He was a founding associate editor of the journal *Language Assessment Quarterly* and has been a series editor of the *Cambridge Studies in Language Testing* (SiLT) series since 2014. He has published widely in language testing and is the co-author of *Learning Oriented Assessment: A systemic approach* with Neil Jones.

Keywords: Assessment, testing, Cambridge, Japan

Interview

This interview took place on Zoom in early March, 2021. Some light editing has been carried out, and clarifications and references have been added where necessary.

Can you give us a brief description of your career in language education and assessment?

I was reflecting the other day that it's 40 years since I started my first teaching job in Sardinia, in Italy, at the University of Cagliari. I taught there for six years, though a lot of my time there was spent doing assessments – I was having to do oral assessments regularly, and to write and deliver exams of English at different levels of a degree program. This brought home to me very soon that, back in 1981, there wasn't very much information about how to do these things. So, where people looked for such information was the existing exams, and what existed back then, in Italy particularly, were the Cambridge exams – the Proficiency and the First Certificate. As I was interested in staying in academia I went back to the UK to do a Master's degree. There I linked up with a group at Reading University doing language testing that included Don Porter, Arthur Hughes[1], and Cyril Weir[2]. And I decided not only to do the Module on language assessment, but also my dissertation on an assessment theme. So, by 1986-7, I was already going in the direction of testing within applied linguistics. Having finished my Master's degree, I came to Japan for the first time. I stayed a couple of years working for Cambridge based in Tokyo and then joined the first Evaluation Unit which was being set up in 1989 with Mike Milanovic in the new EFL department of UCLES, as it was called back then. Over the last 30 years, I've had several roles there, becoming the Director of Research and Validation 20 years ago, which put me in the senior leadership team of the English department. My title is now Director of Research and Thought Leadership, which I've had for the last six years. It's within a much-expanded organization, going from just a few people in the English department to a large research team these days. And that's about where I am now, in the twilight of that career.

A glance at the webpage[3] shows just how many researchers are now working in the validation of Cambridge Assessment English exams…

Yes, it's a massive organization and in a sense the journey of the last 30 years has moved us from a kind of cottage industry with a few experts 'hand-crafting' things to a fully integrated system model. I think one of the successes was to move research into the development phase of exams and the assessment systems and to make it clear that validation is an integrated function – not something that you have done by a few people 'in the shed'! The people that you see on the website who do the research and validation are now involved in designing, developing, and validating over the long term the propositions that we put forward. And that I think has been a great success and where we've shown leadership in Cambridge. Back in the 1980s language testing wasn't really a profession and many exam boards didn't recognize what language testing and evaluation entailed if you wanted to do it to the highest possible standard. Perhaps the Americans were ahead – there was a tradition in psychometrics of setting standards of professional conduct in assessment in the USA. But not in language assessment in the UK, and possibly even less in Europe. I think this has now changed and it was partly precipitated around the time I finished my Master's degree when folk in Cambridge at that time set up the *Cambridge-TOEFL Comparability Study* (see Bachman et al., 1995). That project brought into contrast the different traditions in assessment, and actually flagged up the strengths and weaknesses of both approaches across the Atlantic – a very strong focus on reliability and psychometric principles in the USA, whereas in the UK a very strong focus on what could be called validity, the impact on learning, and the interaction between curriculum and assessment as a design principle dating from the early days of the English exam boards.

Can you tell us about a memorable project you've taken part in?

One of the most interesting projects that I was involved in was the one I joined when I came to Japan in 1987. It put me 'on course' for the rest of my career. I joined a project to tailor the Cambridge exam system, which was a rather underdeveloped set of level-based tests, to meet the needs of Japanese learners. The aim was to introduce two new, level-based tests into what later became the Cambridge Main Suite of exams. This was a multi-partner project funded by the University of Cambridge in collaboration with the British Council and the publisher Kenkyusha that also involved setting up some flagship Cambridge English Schools in Tokyo and Kyoto. I was relatively young back then, starting a new test development project in a new country; I found that extremely interesting and I took away lots of learnings from the experience.

Before I arrived, a project had already been set up to bring Brian Heaton[4] from the UK to work at the University of Tsukuba with Kenji Ohtomo-sensei and his team to revise the Preliminary English Test (or PET) for the Japanese context and also for it to be escalated to-exam status within the Cambridge Exam Suite as a whole. The PET at that time was a minor test, which was developed around 1980 based on the *Threshold Level,* which emerged from the Council of Europe (CoE, 2001) specifications of objectives project, and which was designed more-or-less as a classroom-based formative test.

I came after Brian, and my job was to develop the Pre-PET (what would become the Key English Test, or KET), at the Waystage level from the CoE. In fact, that's when I first started working with the level concepts from the CoE. It was also when I first met the language testers in Japan who later went on to start up and manage the Japan Language Testing Association (JLTA), and who became leaders in the field more generally. I first got to know the STEP exams because the EIKEN model was already a step-by-step, five level system of grades, very much designed for the Japanese system and to cater for what I was told were 'Japanese tastes', but very much unreformed as regards the communicative language teaching shift we'd seen emerging in the 1980s in light of the CoE movement and the *action-oriented approach* to learning, teaching and assessment.

What are some of the exciting things going on at the moment?

In the last decade, what is different is seeing how the promise of technology, both the original Ed-Tech that we've seen in the last 25 years and the new Ed-AI (Ed-Tech + AI), can provide us with the opportunity to develop more transformative, different, and better assessment and learning tasks. A project I'm currently involved in at Cambridge is the University Institute for Automated Language Teaching and Assessment, which brings together my colleagues in the research team with academics in the Computer Lab and the Engineering Department, particularly their speech unit there, together with some other linguists and neuroscientists at the university. We've created a research community in this field, and I think it is quite exciting because we've situated the learning and assessment objectives very clearly at the beginning, rather than working with computer scientists or engineers who come with different perspectives and then try to 'bolt' their ideas onto the constructs and procedures of assessment. I think we are taking a more integrated view, and we can then see how cutting-edge developments in machine learning AI, can be applied to challenges in language learning and assessment from the start. We've created that interface from the beginning, which although not unique, is unusual and is quite exciting. We've got some of the best brains in computer science working with some of the best brains in speech engineering, working with a leading team in learning and assessment. It's very productive intellectually, and likely to lead to solutions which are 'valid' – you could unpack that – designed for the purposes that they are intended from an early stage.

In the last year, I've also been on the advisory board for something called the Institute for Ethical AI in Education, which was set up in the UK under the auspices of Lord Clement-Jones, one of the peers of the realm in the House of Lords. It has brought together three leading thinkers in education in order to come up with recommendations for ethical uses of AI in education. I joined the advisory board about 15 months ago, and the final report is coming out soon[5]. It's intended to provide guidance for educators in general on ethical practices in using AI, but one of the questions I've been asking is, what are the *specific domain-related issues* that crop up in English language learning and assessment? The advantage of assessment is that you capture rich information about people to give them as feedback; the better the information you get about someone, the better the evidence you can provide on what that person can or can't do or is good at in various ways. That is, in a sense, a form of 'surveillance' – you look at people doing things – and the promise of AI is paradoxical because it allows you to surveille people better and capture more data about them. *But do people want to be surveilled? Do they want to be watched in their own homes, for example?* Getting the balance right to take advantage of the AI while at the same time having the checks and balances, the laws, regulations and social practices, is going to be the challenge: it's going to need leadership in our own field of assessment.

What are your thoughts on the current assessment reform proposal in Japan, that is, the use of four-skills tests for university admission?

Over the last 25 years, I've been talking to people working on various grant-funded projects in Japan, including some of the people in the Ministry of Education, Culture, Sports, Science, and Technology (MEXT), concerning the reform of the curriculum. I worked with Professor Koike and his colleagues back in the early 2000s, who were investigating how the learning from the *Common European Framework of Reference* (CEFR; Council of Europe, 2001) could be transferred, adapted, and brought to Japan to introduce a more communicative approach to language education and to benchmark progression through schools from low-level ability to high-level ability to a model like the CEFR. And, as you know, the CEFR itself has been adapted, and the level system used in the revised curriculum. So, we've seen what I would call the *intended uses* of the CEFR, which is a tool for understanding the progression of learning in languages across a school system, as in the Japanese context, with the level system and the learning objectives used as a way of mapping progression across the different cycles.

In Japan's national course of study, MEXT's English curriculum documentation makes it very clear that it is based on a communicative model, that a certain number of hours are required for classroom teaching, that specific materials are to be used, and that specific skills are to be used in classroom activities. But, in the model of alignment that the CEFR promotes, the assessment is required to be in keeping with what's taught and how it's taught. Therefore, if you have moved from a knowledge-based approach to an action-oriented or more communicative approach, then the assessment needs to be based on the same thing as the teaching. The alignment between the curriculum objectives and the outcomes requires the test providers to take that on board. So, when MEXT decided to move to a four-skills approach for university entrance exams, it seemed to be in keeping with the move towards the four-skills curricula in schools (or what now would be multi-modal or integrated skills, covering the productive skills as well as the receptive ones). It was quite exciting to see because it looked like the alignment would be complete. And we all know that without the alignment you get what we would call negative washback, or negative impact, from a test which is out of alignment with the learning. That is, if you teach conversation but only test grammar, increasingly the teaching focuses on the grammar and forgets about the conversation. Unfortunately, the rowing back on the decision to implement the four skills assessments will ultimately slow down the implementation of the curriculum goals because you will not persuade parents, the community at large, teachers and learners, to move to a curriculum based on communication if the exams are based on something else. It's going to be out of alignment for sure.

In Jones and Saville (2016), Learning Oriented Assessment: A systemic approach*, you describe how both classroom and large-scale assessment can work together to support learning. What insights does this approach provide for language educators, curriculum developers, and researchers in Japan?*

I was looking at the book Neil and I put together and I was thinking about what we called a 'systemic approach', and one of the things we have always been worried about is the lack of alignment, or an exam system which prevents you from achieving your goals. And it says here at the end of Chapter 7,

> 'Depending on context there may be several ways of achieving an ecological solution – one where no element of the assessment framework is allowed to subvert the goals of learning'. (Jones & Saville, 2016, p.92)

What I try to point out, both in my thesis and in this book, is that people often only look at one bit of the system and wonder why their reform program doesn't work. They tend to ignore real people, the actual influencers, one or two key people that actually change the way things happen, not 'the government' or 'the ministry'. Or they ignore other aspects of the system for all kinds of reasons, perhaps because it's below their dignity to think about it, but it's those things that are impacting. In the UK, the impact that we were worried about which led us to thinking of things being subversive, were the league tables in schools, where exams are used to judge two things, the learner but also the effectiveness of the teaching and therefore the effectiveness of the school. In the league table model, in order to get a high rating of your school, you need to get high exam results. That meant that schools choose subjects that are easier to get good results in: chemistry is difficult, so 'let's not do chemistry'. That's not an educational goal. So, an educational goal has already been subverted by the league table. Then the teachers say to the students, 'here's what the exam looks like', not three weeks before the exam, but on day one of the course – as students go in to their GSCE course they look at the final exam paper. Well, that's not the intention. What students should look at is what they've got to learn, which is hopefully embedded in the exam and what they'll be tested on, and if they've done the learning it'll be a breeze to pass the exam. What we would call revision or exam prep is perfectly acceptable – I would call that *'seasonal washback'* – we all know when the exam season arrives, there'll be

intensive exam prep – that seems to be normal and acceptable because people shouldn't go into an assessment without knowing what's going to happen. But from day one of the course, possibly two years before students are going to take the exam, they're thinking about it – that kind of '*extensive washback*', way back into the curriculum and the pedagogy and the learning model, is subversive, right? And so, unless we can change it ecologically, not just changing one bit of it, the fundamental changes won't happen.

But it's hard to change people's minds. We started our book with a quote from John Dewey (1933, p. 29-30) to emphasize this:

'It requires troublesome work to undertake the alternation of old beliefs'

The systemic model isn't just about having a great bunch of people in the curriculum department of MEXT, it's still troublesome to alter the beliefs of the people who implement the systems that you are trying to impact. The systemic model is an evolution of the system to bring it into alignment. And old beliefs and new beliefs have to be brought into alignment. It's a change process. The leadership, the top-down needs to impact, but the grass roots also need to be brought on board, the parents and teachers, and the learners, who need to know that what they are taught will actually be assessed. The exams tend to be the authority: if you fail to pass an exam, you might end up on a different path. Everyone knows it and parents behave accordingly: whatever the policy of government, they will try to help their kids be successful. For example, you get negative washback of this kind in China where 'tiger parents' ask teachers in primary schools if their kids can take IELTS[6]. It's a distortion of reality but of the kind that we're talking about.

There is also a problem in the language curriculum as it is currently set out. It is impossible for MOST young people to reach communicative competence beyond a very limited level (A2) just by attending the class hours in the curriculum and doing the associated homework. In other words, treating English as a normal school subject won't work. For a school subject, you have *x* number of hours, say three or four hours a week, plus some homework, alongside all the homework you get for the other subjects. The reason there is a *shadow education system* for language education around the world is because we know you have to connect the school learning with *learning out of school*. This creates more time, more focus, and more consistency in the learning endeavors in order to reach the higher proficiency levels. Young people who become successful learners typically have the opportunity to take advantage of this.

The work of the CEFR-J[7] flagged up where the education system currently is with regard to international levels. And it's low, right, in Japan? It's A1, A2 level. Actually, that's not unusual around the world. The Japan system has exactly the same kind of model of putting English, or whatever language is the target, into a curriculum with *x* numbers of hours a week, with multiple reforms over many years, which lead to minor increments not system-wide change. And I think one of the important things we see coming out of the *Companion Volume to the CEFR* (Council of Europe, 2018) is a focus on learners as social agents, that is they use the language successfully for communication, and it's built into their vision from the start. The old-style model of knowledge-based learning, *'just another subject in the curriculum',* needs to change. That's an attitudinal change towards how languages are learned – and particularly if the language you want to learn is the *lingua franca* of the world that thrives in 'language learning friendly societies'.

Personally, it seems to me you pervert the main educational goal of language learning in the idea that learning is knowledge based. It asks people to waste time and effort on things that ultimately are relatively pointless educationally. It favors that view of the world that you can get on and do it 'in a box'. My abiding early recollections of being in Japan are of how some people are reluctant to do certain things which reveal them in ways that they feel uncomfortable with. And, of course, the concept of the CEFR is all about imperfect progression. Imperfect progression is good. The fact that you speak with all the problems is good. It's not to be shamed; but in some ways in Japan I feel there's a strongly embedded sort of cultural feeling that to reveal your weaknesses is shameful, particularly in ways that could lead to losing face or losing reputation. At the heart of this system-based approach, therefore, are the underlying *cultural norms and mores* of a society – and sometimes they run quite contrary to the learning model for speaking that we're promoting.

So, at the heart of alignment is the ecosystem of a particular context. If you've got this macro-level context, what can you do to succeed at the micro-level? This comes down to the learner interpreting and doing the things she's supposed to do to make progress and be successful. And that's embedded both within a learning model like the CEFR and within the cultural model, and the family model, which influence it. And it's this disconnect or non-convergence that comes with mixed messages or mixed influences which makes it difficult to learn languages in many countries. Because these attitudinal and cultural aspects are not only about the target language, but they also reflect the other languages that are being learned

whether it's the home language, the language of schooling, a wider regional language, or a lingua franca. They bring with them all these cultural and social practices.

In sociolinguistics, I think we're moving towards an understanding that the goal is *social practice*. What we are increasingly trying to do is ask learners to adopt the social practices for communication and not to see the thing they're doing as an academic subject. Of course, if you want to, you can study English literature, and can even do that without really knowing the language. I've met professors in Italy and Japan who don't speak English at all but who know a lot about the literature. That doesn't seem to be ideal, but it can happen. What I'm talking about with the systemic model is to see language as social practice. What you want for Japanese learners is for them to be proficient in their own language and other languages that they need in their lives – and one of them is English. It has been revealed by many studies that the Japanese need English language skills in the current phase of the evolution of this century as much as the Chinese and the Koreans do. So how can you achieve the educational goal of communicative competence without distorting it with the cultural and educational influences, which are from a past era really, about understanding how languages are learned? I would say as a coda to this, get your exam system sorted and a lot of the rest will follow. You have to win the argument about alignment of exams first. The naysayers, the conservative views, need to be put on one side so that this is given a chance, otherwise what you're doing is writing a fiction about language learning in school. You can have a wonderful policy document that shows everything is aligned, but actually the hidden (implicit) curriculum, the pedagogical practices, and the outcomes, will be based on something else. In other words, the implicit curriculum and the social practices override the stated policy.

Concerning the reform, a very recent and somewhat controversial argument is that speaking should not be assessed in high-stakes English exams for university admissions purposes in Japan because students from higher socio-economic status (SES) backgrounds will have an advantage over those from lower SES backgrounds (Butler & Iino, 2021). What are your thoughts on this issue?

I think you're asking the wrong question about SES and speaking. The question is SES full-stop and its impact on learning more generally. The digital divide has proven to be the thing that the Organization for Economic Co-operation and Development (OECD) and other big influencers have flagged up almost from day one of homeschooling during the pandemic (e.g., OECD, 2020). In April last year, when the majority of jurisdictions in the world had moved to emergency remote teaching, for some that meant Zoom all the time, and for others it meant, 'how on earth do we do this because we have no technology'? Some kids can't access anything. If you haven't got the tablet and the broadband, you're completely stuffed. The issue about hybrid models of learning have brought into focus the digital divide, and with it the engineering challenge to ensure that everyone can have broadband, and the economic challenge to make sure everyone can afford it and have a device, or devices if there's more than one child in a household. In some families you've got four kids trying to access one device – so you need four devices not one. Society needs to own this problem and speaking, that is, learning to communicate where speaking is one of a range of skills that you need, should be top of your agenda when it comes to communication, not at the bottom. You're solving the wrong problem if you say let's remove it so that it doesn't become an issue. You should be saying solve the socio-economic access problems and ensure that you can deliver the learning goals for languages as a result.

For language learners who are successful, for example the Scandinavians or the Dutch, research shows that they learn English well in school but what makes it successful is the whole of their society is *language learning friendly*. English exists in society so when they go *out of school* they have opportunities to connect what they've learned in school to something that is really useful in their society. And we need to build on this in Japan. The world has changed for English, especially in the pandemic era, because people have access to English in ways they didn't before through technology. Therefore, you can connect up what goes on in school with what goes on out of school, for conversation classes, for listening comprehension, for authentic task interaction, for game-based learning, for hobby-based learning in a wider sense, for every child. You just have to conquer the digital divide, and to ensure that every child in your country has access to right technology. That would be my answer to the systemic challenge, is to make sure that everyone is digitally connected.

Where does impact research fit in to all of this?

To understand impact, you need to understand what happens in the context where the learning and the use of assessments take place. Impact studies can't be done outside of the context; they need participation from within. You need an understanding of the situational features of the context, and what hypotheses you've got for achieving the intended impact, or the *impact by design*. For instance, if you say you want learners to reach B2 level by the end of high school by 2030, that could be an *impact hypothesis*. If you've designed the system to enable that, (i.e., you've got the CEFR levels in place and over ten years you can raise the level of achievement up one band, so instead of being at B1, a decade later it's B2), now that might be an *impact by design feature*. But how do you know what's happening? This is where washback studies, and

wider impact research come in. Such studies need to feed back into the policy making and pedagogical practice. These studies may well reveal that the intended impact is not being achieved because, for example, the implicit curriculum is still dominating. Why? Well, guess what the pinch point is? The non-aligned exam still exists – the same test that we had before the new curriculum is impacting and preventing the innovation or the evolution of the practices we had in mind. If the assessed objectives are out of alignment, the teacher will have a difficult job because they've got mixed messages: Which voice do they listen to – the voice of the curriculum planner or the voice of the exam? And that becomes the pinch point. Like this, impact by design is a systemic concept and it must be investigated through ongoing impact research.

I think impact research needs to be based on a *theory of impact* and a *theory of action* around how you can achieve the desired impact. How you find out about it is through impact research, and that's not research that's done through the odd impact study, but a program of finding out what happens over time. Therefore, impact is a *longitudinal concept*, it's an iterative concept, and it's a multimodal and mixed-methods concept in terms of research. So, it's like doing a jigsaw puzzle – you find all the pieces and you put them together, but it doesn't fall out of the box and construct itself, you have to piece it together and that takes a long time if you've got a thousand pieces. However, the longitudinal model is not what you normally get in academic studies, in which a researcher conducts a study, publishes the results and moves on to something else. It's more akin to what we in assessment call validation, which is an ongoing requirement to continually build up evidence about what happens. I think many assessment practitioners do not yet hold the view that impact is something we do, just like we do validation. The origins of washback studies can be found largely in academic research but I've tried to put it back into language assessment by saying it's equivalent to validation. Just like working out the reliability of your test every time you administer it. Impact is not something you do once at the time of writing the test, or something you give to an academic to investigate for you once and then you say, well, 'my test has positive washback', which is like saying 'my test is 0.9 reliable'. You can do that, and people do; and that's fine, but it's not job done.

Finally, is there anything exciting happening in 2021 that you wish to let us in on?

I think the most exciting thing emerging from 2021 is a reflection on how the pandemic experience can empower us to think more transformatively about the opportunities we have in front of us to improve multilingual education. The challenge of educational technology and AI is the big one that we're reflecting on. For example, I'm the secretary general of ALTE, and this year we're having our first digital symposium[8]. We will, of course, have more online and hybrid events from now on – as part of the new normal. But our digital symposium is using a virtual reality platform; so instead of taking part in the conference via Zoom and breakout rooms, you'll join a virtual reality conference venue. And you'll take part in a conference that includes sessions, plenaries, networking and coffee breaks. Needless to say, coming to the event will offer challenges. But it will certainly be the first VR conference event I, and possibly many others, will've been to. It only costs 20 euros to sign up, so virtually anyone around the world can join for the first time. So, the impact of the event, its reach, can be enhanced. But it's about balancing the benefits of this technology with other things so that the experience is a rewarding and memorable one for all. That's the challenge we have in front of us.

Thank you, Professor Saville!

Notes

[1] See the classic text, *Testing for Language Teachers*, Hughes & Hughes (2020).
[2] See the recent collection of essays in honor and memory of Professor Weir, *Lessons and Legacy: A Tribute to Professor Cyril J Weir (1950–2018)*, edited by Taylor and Saville (2020).
[3] See https://www.cambridgeenglish.org/research-and-validation/meet-the-team/
[4] See *Writing English Tests,* Heaton (1975)
[5] See https://www.buckingham.ac.uk/research-the-institute-for-ethical-ai-in-education/
[6] International English Language Testing System (IELTS) is a test of English for academic purposes aimed that assesses the language ability of candidates who wish to study in an English-medium tertiary institution. Therefore, for obvious reasons it is not suitable as an assessment for primary/elementary school children.
[7] The CEFR-J is a modified version of the CEFR that is adapted for the Japanese context (see Negishi & Tono, 2016; Tono, 2013). Its development has led directly to revisions to the CEFR itself (see Council of Europe, 2018), particularly regarding lower levels of ability and the distinction between upper and lower abilities within specific levels (e.g., A2.1 and A2.2).
[8] See https://www.alte.org/Digital-Symposium-2021. Editor's note: I attended this conference over three days and thought it was a huge success. Not only was the format a completely new experience, but also the presentations were excellent. A conference to remember!

References

Bachman, L., Davidson, F., Ryan, K., & Choi, I. C. (1995). *An investigation into the comparability of two tests of English as a Foreign Language*. Studies in Language Testing, 1. Cambridge University Press.

Butler, Y. G., & Iino, M. (2021). Fairness in College Entrance Exams in Japan and the Planned Use of External Tests in English, in B. Lateigne, C. Coombe, & J. D. Brown (Eds.), *Challenges in Language Testing Around the World*. Springer.

Council of Europe. (2001). *Common European Framework of Reference for Languages: Learning, teaching, assessment*. Cambridge University Press. https://rm.coe.int/16802fc1bf

Council of Europe. (2018). *Common European Framework of Reference for Languages: Learning, teaching, assessment* (Companion volume with new descriptors). https://rm.coe.int/cefr-companion -volume-with-new-descriptors-2018/1680787989

Dewey, J. (1933). *How we think*. https://www.archive.org/details/howwethink000838mbp

Heaton, J. B. (1975). *Writing English language tests: A practical guide for teachers of English as a second or foreign language.* Longman.

Hughes, A., & Hughes, J. (2020). *Testing for Language Teachers*. Cambridge University Press.

Jones, N., & Saville, N. (2016). *Learning Oriented Assessment: A systemic approach*. Studies in Language Testing, 45. Cambridge University Press.

Negishi, M., & Tono, Y. (2016). An update on the CEFR-J project and its impact on English language education in Japan. In C. Docherty & F. Barker (Eds.), *Language Assessment for Multilingualism: Proceedings of the ALTE Paris Conference, April 2014*. Studies in Language Testing, 44. (pp. 113-133). Cambridge University Press.

OECD. (2020). *OECD policy responses to coronavirus (COVID-19): Combatting COVID-19's effect on children*. http://www.oecd.org/coronavirus/policy-responses/combatting-covid-19-s-effect-on-children-2e1f3b2f/

Taylor, L., & Saville, N. (Eds.) (2020). *Lessons and Legacy: A Tribute to Professor Cyril J Weir (1950–2018)*. Studies in Language Testing, 50. Cambridge University Press. https://www.cambridgeenglish.org/Images/582822-silt-volume-50.pdf

Tono, Y. (Ed.). (2013). *Eigo tōtatsu-do shihyō CEFR-J gaidobukku* [The CEFR-J handbook]. Taishukan Shoten.

Assessing critical thinking in L2: An exploratory study

Sam Reid[1] and Peter Chin[2]
samreid@rikkyo.ac.jp
1. Rikkyo University
2. Waseda University Academic Solutions Corporation
https://doi.org/10.37546/JALTSIG.TEVAL25.1-2

Abstract

Critical thinking (CT) is taking on an increasingly important role in Japanese tertiary education. Teachers tasked with developing CT in a second-language (L2) context may need a way of assessing students' abilities. However, a number of difficulties face L2 students taking a test designed for first-language (L1) speakers. They may be disadvantaged by linguistic and perhaps cultural issues. This study describes an exploratory attempt to make a CT test that can be administered to learners of English and which allows them to display selected elements of CT, specifically analyzing arguments and judging or evaluating. A comparison of L1 and L2 performance in the test showed the results to be comparable. Analysis of two different question topics showed differences in CT skills displayed. Issues with rating accuracy are linked to the format of the test. We argue that this test format is suitable for many students in Japan and elsewhere who have intermediate levels of English.

Keywords: critical thinking, discourse, cognitive load

Critical thinking (CT) may be a feature of tertiary English for Academic Purposes (EAP) courses. Depending on cultural contexts and educational experiences, it may be more or less familiar, and more or less challenging to second language students. It is widely accepted that English as a Foreign Language (EFL) learners require some degree of formal instruction and training in this area (Feng, 2013). This is most especially the case for students going to Anglophone universities. Such institutions often argue that international students do not possess the CT skills necessary for these English-speaking academic cultures (e.g., Fell & Lukianova, 2015; O'Sullivan & Guo, 2011; Shaheen, 2016; Tian & Low, 2011). Furthermore, CT is seen as a requirement to compete in today's global economy (Long, 2004). Despite these academic and financial motivations, there are challenges. In the case of Japan, for example, criticism of students' thinking abilities is a feature of educational discourse (Rear, 2012), and there is growing recognition of "the need for intellectual internationalization and global human resources" (Tsuruta, 2013, p. 147). Considering this, we would agree with Liaw's (2007) view that teachers have a responsibility to help students develop these skills.

CT is a slippery construct. The literature on CT in a first language (L1) differs over issues such as constructs, generalizability, and replication. This uncertainty is heightened when CT is practiced in a second language (L2). Nevertheless, teachers or institutions who make attempts at helping students develop CT need some way of measuring those students' CT ability. This is needed in order to not only assess the current level of their students, but also to track students' progress and measure the effectiveness of CT courses or training within other disciplines that are taught. One option is to use standardized CT tests designed for native English speakers (L1 CT tests). However, as Stroupe (2006) points out, these commercial CT tests may be prohibitively expensive if used on large groups of students. Moreover, as we will argue, using L1 tests for L2 learners may not provide an accurate picture of CT skills, and may be inappropriate for many teaching contexts. An alternative is to convert or translate a test into the students' native language. A drawback to this solution is that it may be expensive and time consuming. Moreover, teachers may also prefer to do a test in English, to serve as more authentic preparation or simulation of study abroad, as well as to simply practice English.

As far as we are aware, no test has been specifically designed to assess CT for L2 students. Therefore, with the aim of ameliorating the disadvantages that L2 students may face, this paper describes an exploratory attempt to make such a test. Our primary concern was whether students taking the test could display equal levels of CT in their L2 as they could in their L1. In addition, we investigated the effect of topic and rating issues. The results indicated that students display similar levels of CT in their L2 and L1. We suggest that the test is flexible and easy to administer, and is particularly suitable for students with intermediate levels of English. We hope it will provide help in the development and guidance of courses meant to foster critical thinking.

Literature review

Although perhaps obvious, it is important to recognize that CT may be more challenging in an L2. A number of studies have looked at the effect of L2 on CT. Davidson and Dunham (1997) used the Ennis-Wier test, in which examinees have to write an evaluation of the arguments in a fictional letter to a newspaper, on tertiary level Japanese students. They found that compared with a control group of 19 students, a treatment group of 17 students who had received instruction in CT did better on the test. Davidson and Dunham's study therefore suggests that it is possible to administer a test designed for L1 to

L2 takers, and that instruction in CT helped them display CT in L2. However, a point to note in this study is that the test conditions did not precisely mirror those for native speakers, as examinees were given twice the standard time to answer the test and were allowed to use dictionaries. In addition, the study does not give any indication of the effect of using CT in L2 rather than L1. Such a comparison is the object of Floyd's (2011) study, in which 55 Chinese students took the Watson-Glaser Critical Thinking Appraisal (Pearson, n.d.). Half of the students took the first half of the test in English and the second half of the test in Chinese, while the other half of the students took the first half of the test in Chinese and the second half in English. Floyd's study used an official licensed translation of the test. The results indicated that displaying CT skills was easier in the L1, a result which was borne out in Floyd's follow-up interviews with participants. There was no time limit and students could use dictionaries, similar to the easing of conditions in the Davidson and Dunham study. A final study addressing this issue is Lun et al. (2010), who were interested in how cultural thinking may affect CT. They administered the Halpern Critical Thinking Assessment using Everyday Situations (Halpern, 2010) and the Watson–Glaser Critical Thinking Appraisal Short Form to students in a New Zealand university. They compared responses from 35 overseas test takers whose L1 was Chinese with those of 24 New Zealand students whose L1 was English and identified as 'New Zealand European'. Results indicated that CT ability was related to general intellectual competence and to English ability, as opposed to cultural thinking styles. In other words, "the difference in critical thinking appears to be more of a linguistic issue rather than a cultural issue" (Lun et al., 2010, p. 613).

There are other studies which use a looser definition of CT, or are not specifically about L2 CT performance, but still shed light on the effect of L2. Luk and Lin (2015) studied a group of Grade 11 students in Hong Kong, and compared what they term 'critical literate talk' in Cantonese and English. Students were tasked with expressing opinions on advertisements, and the definition of CT in this study included generating arguments, evaluating arguments, and making judgments. They found a qualitative difference between the students' ideas expressed in Cantonese and their L2 English, in respect to content and linguistic complexity, and concluded that "The data reveal a wide gap between the students' L1 cognitive maturity and their L2 communicative resources" (2015, p. 70). In terms of written production, Manalo, Watanabe, and Sheppard (2013) investigated university students who wrote about the causes of two disasters, one in their L1 (Japanese) and the other in L2 (English). Their objective was to see if it was harder to be evaluative in Japanese, as Japanese is supposedly less direct in terms of conveying intent or messages. In this study CT was operationalized as students' use of evaluative statements. It is notable again that students were under no time pressure and had received instruction about evaluation. The results showed that students produced more evaluative sentences, evaluative sentences about causes, and evaluative sentences with support when writing in Japanese compared to English. The effect of CT in L2 is shown by "significant correlations between the students' TOEIC scores and their production of evaluative sentences in English (their L2) – but not in Japanese (their L1)" (2013, p. 2971). In other words, their results suggest that although CT is not a linguistic skill, its clear expression requires linguistic ability.

A study by Kaupp et al. (2014) gave three different CT tests to first year students in a Canadian university in an attempt to form a more comprehensive measure of students' CT. Their primary purpose was to assess students' CT development over the course, but as some of the students had English as their L2, they commented on this group when discussing their results. They found that among the three standardized CT tests used – the Cornell Critical Thinking Test: Level Z (Ennis et al., 1985), the International Critical Thinking Essay Test (Paul & Elder, 2010), and the Collegiate Learning Assessment (Council for Aid to Education, n.d.) – only the results from the latter showed significantly lower performance by the English L2 group. Another paper which included a similar analysis of a group of English L2 university students was Facione (1990b), which applied the California Critical Thinking Skills Test (College Level) (Facione, 1990c) to 1,196 students at an American university. Non-native speakers comprised 19% of the sample, and Facione found statistically significant differences between native English speakers and non-native English speakers, who scored lower. His conclusion is unequivocal: "That there is no significant difference from pretest to posttest for non-native English speakers indicates that the CTST instrument is not appropriate for the assessment of college students who are not native English speakers" (1990b, p. 12). The research described so far thus underscores how one must be careful not to mistake a lack of linguistic ability for a lack of CT ability.

Clearly, much of the literature suggests CT is more difficult in L2, so the next issue is why this should be the case. The starting point is the central role of language in CT. According to Moon (2008), although the importance of language differs between CT activities, "it must be seen as extremely important in any critical thinking in the manner that the communication of the thinking is conveyed, distorted, precise or not precise, clear or not clear, subject to manipulation, filled with assumptions, and so on" (p. 73). Similarly, Kobrin et al. (2016) emphasize how language is crucial for both understanding and as a tool for expressing CT. Such views are supported by Takano and Noda (1993), who found that performance in a thinking task declined when a concurrent linguistic task had to be performed in a foreign language. These difficulties are clearly a factor when students take a CT test designed for native speakers. For instance, the Watson-Glaser has been criticized for its unclear instructions and confusing terminology (Possin, 2014). Tellingly, as Kennedy et al. (1991, as cited in Lai, 2011) note, commercially available US CT tests are not designed for students below the fourth-grade level, so tests assume a certain level of linguistic competence that L2 students may not possess. In surveying the constructs which are

assessed by CT tests, Kobrin et al. (2016) note that "Despite differences in the specific knowledge, skills, and abilities measured across critical thinking tests . . ., they all require some verbal ability" (p. 4). This is a particular issue for CT tests which require writing passages. L2 examinees without an advanced level of L2 fluency are bound to make lexical and syntactic errors in their writing and thus may not convey their intended meaning. A rater might disregard such an answer as unacceptable or unclear. Furthermore, L2 examinees may simply decide not to write certain viewpoints because they feel they lack the lexical knowledge to properly explain them in L2. Any attempt to test CT in L2 should take these potential linguistic obstacles into account.

In addition to linguistic knowledge, it has been suggested that differing factual and even cultural knowledge may hamper students. One of the contested points in CT research is about whether CT skills are specific to content areas, or are universal (Moore, 2004). According to Lai (2011), most CT researchers believe background knowledge is important, being "a necessary, though not sufficient, condition for enabling critical thought within a given subject" (p. 42). Norris (1985), for example, argues that successful application of CT requires "among other things, a knowledge of the subject matter, experience in the area in question, and good judgment" (p. 44). To give examples of potential problems with lack of background knowledge, on the California Critical Thinking Skills Test the final four questions refer to a story of a white supremacist and an accompanying scenario in an American school setting, where issues of poverty and race relations arise. In a similar vein, the Ennis-Weir test involves analysis of overnight parking problems. Although this issue is less culturally specific, levels of car ownership and the importance of parking restrictions are not the same in all societies. An important study of whether content familiarity plays a role in critical thinking in relation to L2 writing is Stapleton (2001). In his study of Japanese university students, half the students wrote about rice importation, and half about gun control in the U.S. He found a broader range of arguments and evidence deployed for the familiar topic of rice importation, greater levels of abstraction about the topic, and more references to other viewpoints on the issue. He explains how it is hard to go beyond the literal ideas in the prompt if you do not have background knowledge to tie these ideas to, as wider schema facilitate deeper abstraction about a topic (p. 530). A related study is He and Shi (2012), who tested the effect of topic knowledge on the writing performance of 50 Canadian ESL students with varying degrees of English proficiency. They found that writing performance was better on the general topic of university studies than the specific topic of federal politics. Although this study was not focused on CT, the differences were in "poor idea quality, insufficient idea development, implicit position taking, and weak conclusions" (p. 460), which fall under the scope of CT. Another factor to consider in assessing CT, therefore, is topic choice.

Finally, knowledge may not be a purely factual construct, and may extend to a way, or manner, of thinking. This has important implications for L2 CT, particularly for multiple-choice test formats. Both Ennis (1993, p. 181) and Taube (1995, p. 15) point to how test takers with different assumptions and background beliefs to the test authors' may follow logical lines of reasoning, but will not receive credit for selecting an 'incorrect' answer in tests with a multiple-choice format, where usually no opportunity is given for students to explain the logic behind their selected items. The more distant a student's cultural background, the more likely this becomes. For instance, Fawkes et al. (2005) identified answer choices for the California Critical Thinking Skills Test that potentially have multiple interpretations, thus affecting what can be considered a 'correct' answer. In addition to issues specific to forced choice tests, the vexing topic of 'cultural influence' is relevant to CT more generally. On one side of the debate are Ramanathan and Kaplan (1996), who caution that CT tests examine cultural knowledge that L2 learners may not share, and Atkinson (1997), for whom CT is a social practice better described as cultural thinking. The riposte to these ideas is characterised among others by Davidson (1997), who argues it is more accurate to say that CT is tolerated to different degrees in different spheres of cultures, and Paton (2005), who believes the reasons for student difficulties are lack of practice and topic knowledge rather than thinking styles. With this in mind, raters may need to be aware of such potential differences.

In sum, the difficulties faced by students are neatly summarised by Bali (2015) as "their cultural capital and exposure to critical thinking before college; their exposure to pedagogies that promote critical thinking before college; and their linguistic ability, which impacts their ability to read/write critically" (p. 327). With these things in mind, any attempt to measure CT in a second language has to allow for lower linguistic ability, allow for differences in background assumptions, and allow for differing topic knowledge. This is relevant in the next section, which describes the test format.

Test Format

A number of basic factors were considered important in the context of designing a CT test for an L2 situation. First is the answer requirement. Although a test with a forced choice format can make implementation and rating manageable, as the literature review detailed, this may not be ideal for testing CT. In a synthesis of research on critical thinking, Norris (1985) stated that ideally CT testing requires that takers be productive, not just choose correct options or avoid errors. A second important point is explained by Stroupe (2006), who stressed how assessment of learners' CT in L2 situations must be level appropriate. Lai (2011) advised that "In constructing assessments of critical thinking, educators should use open-ended

tasks, real-world or 'authentic' problem contexts" (p. 42). The final principle is articulated by Facione (1990a), who listed the constructs which should not advantage nor disadvantage students doing a CT test, and among these the important points were reading ability, background knowledge, and culture (p. 32).

A pilot test was carried out in which students were given 20 minutes to write critical responses to the statement "Learning English is necessary for success in today's world". Students had to explain in as much detail as possible why this might not be true. The pilot showed that students required more specific instructions on how to answer, and that students tended to stop writing after about 10 minutes. Based on this, it was decided to give students 10 minutes to write, to provide examples of how to answer, and also to provide two different statements to critically respond to, in order to compare the ease of responding to different topics. Therefore, in the version of the test we trialed, students were given two statements:

- "Learning English is necessary for success in today's world." (henceforth "Learning English")
- "All endangered animals should be saved." (henceforth "Endangered Animals")

These were designed with attention to vocabulary and length, to reduce anything lexically and syntactically challenging to L2 examinees of minimum low-intermediate level. The two topics were chosen based on the authors' attempt to limit possible advantages or disadvantages for students with or without specialist background knowledge. It was presumed that students of diverse cultures would have had sufficient exposure to both issues to be able to articulate views on them. In this way, it was hoped the test was both level appropriate and fair in terms of contextual knowledge. The test was 'productive' in that students had to write as many ideas challenging the statements as possible. Both statements were purposefully vague and easy to challenge, to encourage as wide a variety of responses as possible.

In terms of operationalizing CT for assessment, there is disagreement in the literature as to the scope of what should be tested and the processes that constitute CT. However, as Liaw (2007) argued, there is little essential difference in the various critical thinking definitions. Perhaps the most influential definition of what constitutes CT is Facione's (1990a) consensus statement. This lists the cognitive skills of interpretation, analysis, evaluation, inference, explanation and self-regulation; and the dispositions of critical thinkers, which include open-mindedness regarding divergent world views, flexibility in considering alternatives and opinions, understanding of the opinions of other people, fair-mindedness in appraising reasoning, honesty in facing one's own biases, prejudices, stereotypes, egocentric or sociocentric tendencies, and prudence in suspending, making or altering judgments. This is fairly broad, and for our purposes critical responses were classed as responses which questioned the validity of the original statement, such as counter-arguments, questions about the logic of the statement, or combinations of both. To decide if a response was acceptable, four underlying critical thinking skills were possible:

1. *Seeking clarity* (Is the concept clear? Does any language need to be clarified?)
2. *Challenging the logic* (How true is the statement? Is there any doubt as to its possibility?)
3. *Presenting an alternative viewpoint* (Are there possible negative consequences to consider? Are there more important issues regarding any point in the statement?)
4. *Challenging an assumption* (What are the statement's underlying assumptions? Are these assumptions valid?)

Any response that fulfilled these criteria was classed as acceptable (examples are provided in Appendix A). The more responses the examinees could produce to question the validity of the statements, the better their level of CT. This test therefore focuses on two elements of CT: analyzing arguments, and judging or evaluating. These fit McPeck's (1981, p. 8) definition of CT as "reflective skepticism".

The literature review identified difficulties with L2 and cultural background knowledge as factors which may impact the display of CT. Therefore, when analyzing test performance, we were interested in whether students were disadvantaged by writing in L2. We also wanted to compare the two topics to determine any effect of background knowledge. Finally, as this is an exploratory study, we wanted to see how accurately our operationalization of CT could be judged. With this in mind, the three research questions were as follows:

1. Can participants display evidence of equal CT skills in their L2 as well as they can in their L1?
2. How do the responses produced by the students for each of the two statements compare in terms of number and acceptability?
3. Can the L2 CT test be accurately rated?

Test Implementation

The test was administered at a private Japanese university. The participants comprised 138 students, of whom 102 identified their native language as Japanese, and 36 as Chinese. The students were enrolled in an elective course that focused on critically reading English texts, such as articles and short stories, and discussing their analyses of the readings. The course is recommended for students with a minimum TOEIC 600 or TOEFL iBT 64. However, students were not required to provide proof of TOEIC or TOEFL scores, so could conceivably have been lower, and we did not have measurements of language proficiency. Students took the CT test in the first lesson of the course. All students signed a consent form.

Students had to write as many critical responses to each statement as possible in 10 minutes. Critical responses were single sentences (not paragraphs). For the first statement, students were instructed to write their critical responses in English, and for the second statement in their native language. To illustrate the task requirements and to show students that the L2 linguistic demands on this test are presumably within their range, prior to the start of the test, students were given an example statement ("Dogs make the best pets") and a list of critical responses to the statement (see Appendix B). While writing, students were neither allowed to speak nor use a dictionary.

Two versions of the tests were administered: Practice A and Practice B. In Practice A, students responded to "Learning English" in English, and responded to "Endangered Animals" in their L1. In Practice B students responded to "Endangered Animals" in English and "Learning English" in their L1. Classes were assigned at random to do either Practice A or Practice B. In total, 65 students completed Practice A and 73 completed Practice B.

To ensure accurate rating of the responses written in Japanese and Chinese, we enlisted translations from native speakers. Two raters (the authors of this paper) checked all responses independently and marked each response as acceptable or not. It was agreed in advance that in the case of poor grammar or vocabulary we would give students the benefit of the doubt, following Paul and Elder's (1996, as cited in Stroupe, 2006) suggestion that when assessing CT intellectual standards should be concerned with reasoning over quality of writing. Raters then compared answers. Upon disagreement over acceptability, raters discussed the interpretation of the categories and decided upon a final judgment of acceptable or unacceptable. Examples of acceptable and unacceptable responses can be seen in Appendix A.

Results

The first research question was posed to determine whether participants could display the same level of CT in their first and second languages. Table 1 compares the number of acceptable responses for each statement in participants' L1 and L2. There were more acceptable responses in L1 than in L2 for both statements. The difference was 0.20 more responses in L1 for "Learning English", and 0.31 more responses for "Endangered Animals". The participants in this study included both Japanese and Chinese students, which offered a chance to compare responses by participants' L1. Table 2 breaks down acceptable responses in L1 and L2 by Japanese and Chinese participants. Chinese participants produced a marginally higher mean number of combined L1 and L2 responses for each statement. The difference between the number of L1 and L2 responses was higher among Japanese participants than among Chinese participants. Notably, in the case of Chinese participants responding to "Endangered Animals" there were 0.04 more responses in L2 than L1.

Table 1

Mean acceptable responses for each statement in L1 and L2

Statement	Group	M	SD
Learning English	L1	4.55	2.14
	L2	4.75	2.06
	L1 & L2	4.66	2.09
Endangered Animals	L1	5.97	2.70
	L2	6.26	3.22
	L1 & L2	6.11	2.95

Note: Practice A: n =65; Practice B: n = 73

Table 2

Mean acceptable responses for each statement by participant L1

Participant L1	Statement	Group	M	SD
Japanese	Learning English	L1	4.71	2.15
		L2	4.43	2.15
		L1 & L2	4.6	2.14
	Endangered Animals	L1	6.35	3.45
		L2	5.94	2.64
		L1 & L2	6.10	2.97
Chinese	Learning English	L1	5.00	1.48
		L2	4.76	2.15
		L1 & L2	4.83	1.95
	Endangered Animals	L1	6.12	2.88
		L2	6.18	3.16
		L1 & L2	6.14	2.92

Another possible indication of whether participants could display equal CT in L1 and L2 was the number of responses written in each language that were rated as acceptable. If expressing ideas is more difficult in L2, there may be fewer acceptable responses in L2 compared with L1. Table 3 shows the percentage of responses graded as acceptable, and whether responses were in L1 or L2. For both statements, more responses were graded as acceptable in L1. Taking L1 and L2 together, 14.74% more responses were rated as acceptable for "Endangered Animals" than "Learning English".

Table 3

Percentages of responses graded acceptable for each statement by participant L1

Statement	Group	Acceptable (%)
Learning English	L1	71.99
	L2	64.77
	L1 & L2	68.48
Endangered Animals	L1	85.32
	L2	81.94
	L1 & L2	83.22

Note: Practice A: n =65; Practice B: n = 73

The second research question was posed to compare the number and acceptability of responses to each statement. Table 1, above, shows that the mean responses to "Learning English" was 4.66 and the mean responses to "Endangered Animals" was 6.11, and so 1.45 greater for "Endangered Animals". In order to add detail to this, Table 4 shows the mean number of acceptable responses per participant, as well as the overall number of responses to each statement, and the variety of different ideas given for each statement (in other words, different possible responses). As well as having more overall responses, "Endangered Animals" had a greater variety of responses in terms of content. This shows that there were more ideas produced in response to "Endangered Animals".

Table 4

Number of acceptable responses and variety of acceptable ideas for each statement

	Learning English	Endangered Animals
Overall number of responses	643	843
Variety of ideas	113	135

Figures 1 and 2 are box plots comparing the mean number of responses to each statement broken down into nationality and L1/L2. They show how "Endangered Animals" was easier to respond to compared with "Learning English", as indicated by inter-quartile spreads as well as bottom and top whisker lengths. In Figure 1, among the Japanese L1 participants, boxplot spread, skew, and median value were identical regardless of L1 or L2 use. In Figure 2, there were differences when it came to responding to "Endangered Animals", with a wider range of responses in L2. The spread of responses among the Chinese participants were more compact compared to the Japanese participants.

The third research question concerned whether the test could be accurately rated. There were two stages to the grading process. To begin with, graders separately checked all responses and decided upon acceptability. Second, graders came together to compare results and discuss cases of disagreement over acceptability to decide these cases together. When raters compared together after the first separate check, inter-rater agreement of acceptable responses to "Learning English" was 84.96%, and agreement of acceptable responses to 'Endangered Animals" was higher at 92.59%. For both statements combined it was 89.11%.

Figure 1

Acceptable responses to "Learning English" by L1 and L2

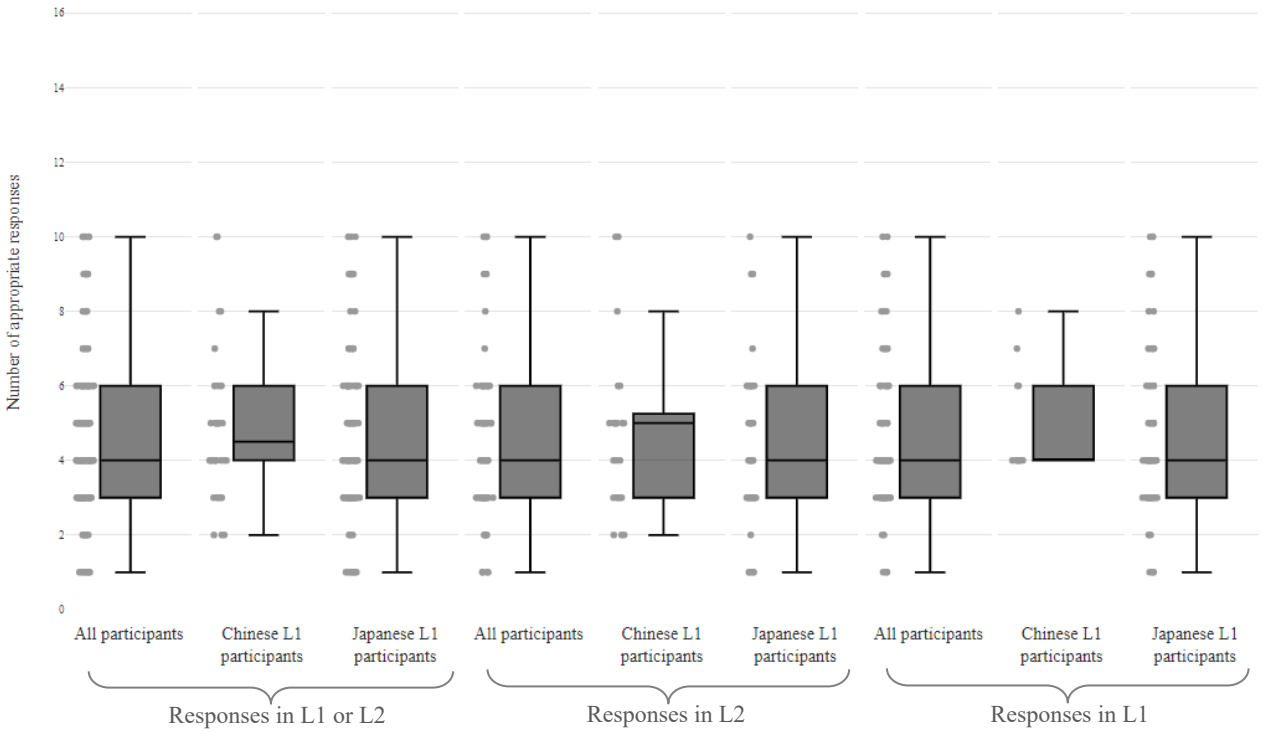
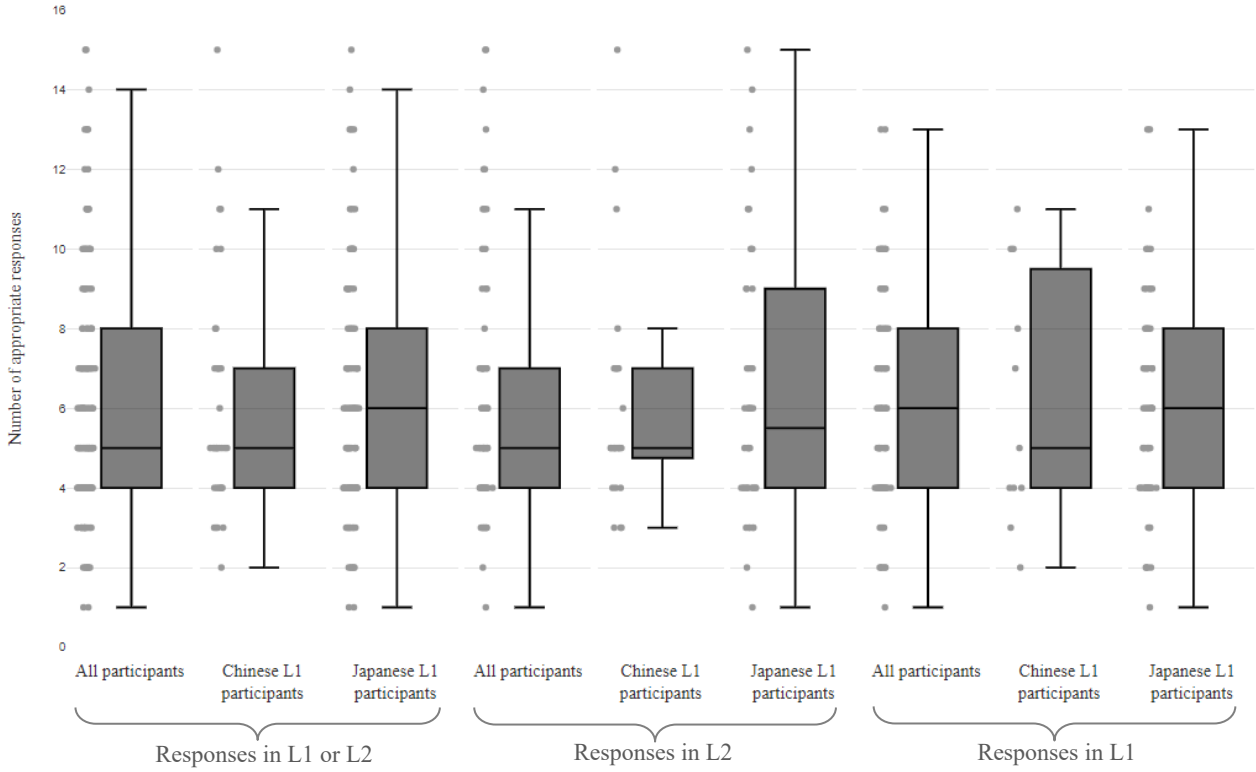


Figure 2

Acceptable responses to "Endangered Animals" by L1 and L2



Discussion

To recap, the research questions concerned the level of CT participants could display in L1 and L2, the comparative difficulty of the two statements, and the accuracy of rating.

Research Question 1: Can participants display evidence of equal CT skills in their L2 as well as they can in their L1?

Participants performed slightly better in their L1 than their L2: The difference in mean acceptable L2 responses were 4.55 to 4.75 for "Learning English", and 5.97 to 6.26 L1 for "Endangered Animals". Therefore, the goal of creating a test which did not disadvantage L2 participants was not completely successful, even though the differences between L1 and L2 responses do not appear extreme. Overall, then, this study supports the idea noted elsewhere that displaying CT skills is more difficult in a L2 due to linguistic difficulties, rather than deficient CT ability. However, the fact that Chinese students responding to "Endangered Animals" produced a slightly higher mean of acceptable L2 responses than in L1 suggests that language may not always be an inhibiting factor. The study showed slightly higher CT levels for the Chinese students. This could be because students who have the aptitude and resources to study abroad (as is the case with Chinese students studying in Japan) possibly have higher academic levels and/or language proficiency.

As well as a higher mean of acceptable responses, the results showed a slightly higher percentage of responses were rated as acceptable in L1 than L2. Interestingly, however, language comprehensibility was not an issue. All but three of the responses written in L2 were comprehensible for the raters. Answers were overwhelmingly deemed unacceptable through weak CT rather than lack of comprehensibility. As there were 992 total responses in English, to have only three responses (0.30%) determined to be incomprehensible due to issues with L2 lends support to the format of the test being appropriate for L2 students at a proficiency level of around TOEIC 600 and over. It is difficult to ascertain the reason for this greater response acceptability in L1. One possibility is that participants felt more confident expressing ideas in L1 and hesitated to express ideas in L2.

A drawback of this test format is the limited operationalization of CT in the test. In making a test that is suitable for students who cannot produce longer written passages certain compromises were necessary, one of which was that the test focused on deconstructing rather than constructing arguments. It does not measure other possible CT aspects, such as the ability to argue for a position, using supporting reasons and examples, judging evidence, or deciding on a course of action, all of which have been included in definitions of CT. Other tests that have been found to disadvantage L2 participants more than this one may measure a more comprehensive operationalization of CT. Another drawback is that L2 ability has been identified as a factor in display of CT in L2, but we did not take L2 ability into account. Regrettably, we did not have a standardized test measure to factor in, such as a TOEIC or TOEFL score. Students on these courses were expected to have an English level of intermediate and above, but we lacked the means to differentiate students based on English levels.

Research Question 2: How do the responses produced by the students for each of the two statements compare in terms of number and acceptability of responses?

The results clearly show that "Endangered Animals" was easier to respond to than "Learning English". Mean responses were higher, the percentage of responses rated as acceptable was higher, and the variety of ideas rated as acceptable was higher. The higher inter-rater agreement about responses to "Endangered Animals" may also suggest that this was easier to respond to. "Learning English" may be a more complex issue in terms of critical responses for a number of reasons. Perhaps students' first-hand experience with this topic made responding critically to long-held assumptions about it more formidable, or contributed to students drawing more on illogical or fallacious ideas than when responding to "Endangered Animals". Regardless of the reason, the present study lends support to the finding that topic affects CT display. The fact that the range of responses differed more for "Endangered Animals" might indicate that the more open to criticism a statement is, the more the use of L1 or L2 can impact performance. If this is true, it would affect the degree to which the test is able to differentiate high from low performers, both regarding CT ability (number of acceptable responses) and strength of L2 ability (in comparison to L1 results).

Further research would help clarify which topics may be most appropriate for general or particular kinds of English L2 students. For instance, other statements could be "Famous people are good role models because they are successful" or "It is important to protect the natural environment for future generations", as it would be assumed most cultures have celebrities and protecting the environment is a universally debated topic. In addition, future research could involve interviews with participants to ask whether one topic was more difficult than the other and why.

Research Question 3: Can the L2 CT test be accurately rated?

With respect to rating, we consider 89.11% initial agreement on how to classify responses an acceptable level, considering the open-ended format of the test, potential variability in interpretations of criteria, and differences in the cross-cultural backgrounds of those involved in the test (both students' and raters'). All differences were resolved through rater discussion, which served to further refine the criteria. Rating difficulties is a complex topic that merits more discussion.

One source of disagreement was about the scope of the categories. For instance, two responses to "Learning English" were *Might produce a world centering only on U.S.* and *Lead to a lack of study of native language*. One rater considered these to be unacceptable because they are not problems with the logic or feasibility of the statement itself, but arguments against "Learning English". However, after discussion, it was decided to include criticisms based on unintended or undesirable consequences. A further example concerns the following responses to "Endangered Animals": *It is decided by God* and *Eating whale is a Japanese tradition.* Disagreement centered on whether to allow for possible adherence to religion or traditions (which are not CT in the sense of analytical logic). For instance, in some cultures, it is believed that nature is controlled by God, and that upholding tradition is regarded as more important than protecting endangered animals. These responses were eventually deemed acceptable in consideration of what might have been positions based on the students' cultural or personal beliefs.

The second type of problem was with the level of detail required for an acceptable response. This area is arguably more complex. To illustrate, the following responses were frequently given for both statements: *It takes too much time* and *It costs too much money.* These answers seem to imply the following: "The time and money which would be spent on these endeavors would be better used for other purposes". Again, the issue comes up of differing background beliefs. We assumed that students felt that this was shared context that did not require further explanation, and decided to accept such responses as acceptable. However, the inferences involved in other responses were less clear. For example, two responses to "Learning English" were *Nationality is more important* and *It reduces your chances to learn other languages.* Similarly, responses to "Endangered Animals" were *Humans are animals too, and it is selfish to regard humans as different from other animals* and *People may use them to make money.* It was decided that such responses could have been acceptable had the students explained their thoughts further. However, at some point the potential reasons were numerous or not immediately obvious, and it was at this point that an answer became unacceptable.

This issue of category boundaries may be resolvable with clearer pre-rating guidelines. For example, a list of standardized acceptable and unacceptable answers given to students before the test would be useful. However, the scope of acceptable answers is a more serious issue. The rating examples highlight an intrinsic difficulty with the format of the test (and perhaps any CT test based on assessing production), which is the subjective nature of the assessment criteria. There is a need for a cut-off point between what is inferable (and thus acceptable), and what requires further explanation (and thus is not acceptable).

Another limitation of this test format is the ambiguity of deciding the cut off point for what constitutes enough support or explanation for an idea. We did not pre-discuss the implementation of categories because we wanted to see what issues would become apparent. The difficulty in this study with identifying CT aligns with what has been noted by others. For instance, in rating L2 essays for evidence of CT skills Stapleton (2001) commented that agreeing on categorization was difficult, and agreeing on what was acceptable or not acceptable was also difficult. He also noted that while there is discussion of the idea of critical thinking itself, there are few criteria or scoring guides. In other words, it is easier to provide an abstract definition of CT than to delineate a cut-off point between a concrete phrasing that is and is not sufficiently critical. On a similar note, Possin (2014) highlights the difficulty in how judgments about the acceptability of conclusions may be different between test takers and test raters because of differing background beliefs. Finally, Norris (1989) stresses that while any answer key represents a test maker's judgment of what is acceptable, the "test maker must take into account … background empirical beliefs and political and religious ideologies that reasonably could be expected to be held by examinees, and assumptions that examinees would likely make" (p. 23). These issues came to the fore when rating, and it appears cultural differences in what constitutes shared knowledge and CT expectations contributed to some statements being deemed unacceptable. The ongoing debate over definitions of CT suggest that it is also an issue in L1 assessment, not just L2 assessment.

Conclusions

With this CT test format, participants made more responses in their L1 than in their L2. Also, the topic that participants responded to was an important factor in how much CT they could display. Finally, issues for grading were not related to understanding student responses in terms of the linguistic content, but about the boundaries of what constitutes CT. Overall, our study supports previous findings concerning the increased difficulty of CT tasks when performed in the L2. However, we do not feel that the small differences observed between the L1 and L2 performance negate the usefulness of the test. The test format is quick to implement and is easily adaptable in terms of the statements to respond to. Such a format may point the way to an appropriate testing instrument for the many instructors in Japan and other countries with students who do not have a suitable level of English for an L1 test, or who have not been trained in formal writing in the L2. We encourage others to modify and refine this format.

Acknowledgements

The authors are very grateful for the detailed feedback provided by the reviewers and editor.

References

Atkinson, D. (1997). A critical approach to critical thinking in TESOL. *TESOL Quarterly, 31*(1), 71-94. https://doi.org/10.2307/3587975

Bali, M. (2015). Critical thinking through a multicultural lens: Cultural challenges of teaching critical thinking. In M. Davies, and R. Barnett (Eds.), *The Palgrave Handbook of Critical Thinking in Higher Education* (pp. 317-334). Macmillan.

Council for Aid to Education. (n.d.). *Collegiate Learning Assessment*. https://cae.org/

Davidson, B. W. (1997). Comments on Dwight Atkinson's "A critical approach to critical thinking in TESOL": A case for critical thinking in the English language classroom. *TESOL Quarterly, 32*(1), 119-123. https://focionline.files.wordpress.com/2014/05/atkinson-comment-1.pdf

Davidson, B. W., & Dunham, R. A. (1997). Assessing EFL student progress in critical thinking with the Ennis-Weir Critical Thinking Essay Test. *JALT Journal, 19*(1), 43-57. http://jalt-publications.org/jj/articles/2704-assessing-efl-student-progress-critical-thinking-ennis-weir-critical-thinking-essay

Ennis, R. H., Millman, J., & Tomko, T. N. (1985). *Cornell Critical Thinking Tests Level X & Level Z: Manual.* Midwest Publications.

Ennis, R. H. (1993). Critical Thinking Assessment. *Theory into Practice, 32*(3), 179-186.

https://doi.org/10.1080/00405849309543594

Facione, P. A. (1990a). *Critical thinking: A statement of expert consensus for purposes of educational assessment and instruction*. https://files.eric.ed.gov/fulltext/ED315423.pdf

Facione, P. A. (1990b). The California Critical Thinking Skills Test-College Level. Technical Report #2. Factors Predictive of CT Skills. California Academic Press.

Facione P. A. (1990c). *California Critical Thinking Skills Test: College Level.* California Academic Press.

Fawkes, D., O'Meara, B., Weber, D., & Flage, D. (2005). Examining the exam: A critical look at The California Critical Thinking Skills Test. *Science & Education, 1*(4), 117-135. https://doi.org/10.1007/s11191-005-6181-4

Fell, E. V., & Lukianova, N. (2015). British Universities: International students' alleged lack of critical thinking. *Procedia – Social and Behavioural Science, 215*(8), 2-8. https://doi.org/10.1016/j.sbspro.2015.11.565

Feng, Z. (2013). Using teacher questions to enhance EFL students' critical thinking ability. *Journal of Curriculum and Teaching, 2*(20), 147-153. https://doi.org/10.5430/jct.v2n2p147

Floyd. C. B. (2011). Critical thinking in a second language. *Higher Education Research & Development, 30*(3), 289-302. https://doi.org/10.1080/07294360.2010.501076

Halpern, D. F. (2010). *Halpern Critical Thinking Assessment.* Schuhfried.

He, L., & Shi, L. (2012). Topical knowledge and ESL writing. *Language Testing, 29*(3), 443-464. https://doi.org/10.1177/0265532212436659

Kaupp, J., Frank, B., & Chen, A. (2014). *Evaluating critical thinking and problem solving in large classes: Model eliciting activities for critical thinking development*. Higher Education Quality Council of Ontario. http://www.heqco.ca/SiteCollectionDocuments/Formatted%20Queen%27s_Frank.pdf

Kobrin, J. L., Sato, E., Lai, E., & Weegar, J. (2016, April 9-11). *Examination of the constructs assessed by published tests of critical thinking* [Paper Presentation]. Annual Meeting of the National Council on Measurement in Education, Washington, D.C.

Lai, E. R. (2011). *Critical thinking: A literature review*. Pearson. http://images.pearsonassessments.com/images/tmrs/CriticalThinkingReviewFINAL.pdf

Liaw, M. (2007). Content-based reading and writing for critical thinking skills in an EFL context. *English Teaching & Learning, 31*(2), 45-87. https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.580.288&rep=rep1&type=pdf

Long, C. C. (2004). *Teaching Critical Thinking in Asian EFL Contexts: Theoretical issues and Practical Applications*. www.paaljapan.org/resources/proceedings/PAAL8/pdf/pdf022.pdf
http://paaljapan.org/resources/proceedings/PAAL8/pdf/pdf022.pdf

Luk, J., & Lin, A. (2015). Voices without words: Doing critical literate talk in English as a second language. *TESOL Quarterly, 49*(1), 67-91. https://doi.org/10.1002/tesq.161

Lun, V. M., Fischer, R., & Ward, C. (2010). Exploring cultural differences in critical thinking: Is it about my thinking style or the language I speak? *Learning and Individual Differences, 20*, 604–616. https://doi.org/10.1016/j.lindif.2010.07.001

Manalo, E., Watanabe, K., & Sheppard, C. (2013). Do Language Structure or Language Proficiency Affect Critical Evaluation? *Proceedings of the Annual Meeting of the Cognitive Science Society, 35*(35), 1069-7977.

McPeck, J. E. (1981). *Critical thinking and education.* St. Martin's Press.

Moon, J. (2008). *Critical thinking: An exploration of theory and practice.* Routledge.

Moore, T. (2004). The critical thinking debate: How general are general thinking skills? *Higher Education Research & Development, 23*(1), 3-18. https://doi.org/10.1080/0729436032000168469

Norris. S. P. (1985). Synthesis of research on critical thinking. *Educational Leadership*, 40-45. http://www.ascd.org/ASCD/pdf/journals/ed_lead/el_198505_norris.pdf

Norris, S. P. (1989). Can we test validity for critical thinking? *Educational Researcher, 18*(9), 21-26. https://doi.org/10.3102/0013189X018009021

O' Sullivan, M. W., & Guo, L. (2011). Critical thinking and Chinese international students: An East-West dialogue. *Journal of Contemporary Issues in Education, 5*(2), 53-73. http://dx.doi.org/10.20355/C5NK5Z

Paton, M. (2005). Is critical analysis foreign to Chinese students? In E. Manalo. & G. Wong-Toi (Eds.), *Communication skills in university education: The international dimension* (pp. 1–11). Pearson Education.

Paul, R., & Elder, L. (2010). *International Critical Thinking Test.* Foundation for Critical Thinking.

Pearson. (n.d.). *Watson Glaser Critical Thinking Appraisal*. https://www.talentlens.co.uk/product/watson-glaser/

Possin, K. (2014). Critique of the Watson-Glaser Critical Thinking Appraisal Test: The more you know, the lower your score. *Informal Logic, 34*(4), 393-416. https://doi.org/10.22329/il.v34i4.4141

Ramanathan, V., & Kaplan, R. B. (1996). Some problematic "channels" in the teaching of critical thinking in current LI composition textbooks: Implications for L2 student-writers. *Issues in Applied Linguistics, 7*(2), 225-249.

Rear, D. (2012). The dilemma of critical thinking: conformism and non-conformism in Japanese education policy. In T. Isles & P. Matanle (Eds.), *Researching Twenty-First Century Japan: New Perspectives for the Electronic Age* (pp. 119 – 137). Lexington Books.

Shaheen, N. (2016). International students' critical thinking–related problem areas: UK university teachers' perspectives. *Journal of Research in International Education, 15*(1), 18-31. https://doi.org/10.1177/1475240916635895

Stapleton, P. (2001). Assessing critical thinking in the writing of Japanese university students: Insights about assumptions and content familiarity. *Written Communication, 18*(4), 506-548. https://doi.org/10.1177/0741088301018004004

Stroupe, R. R. (2006). Integrating critical thinking throughout ESL curricula. *TESL Reporter, 39*(2), 42-61.

Takano, Y., & Noda, A. (1993). A temporary decline of thinking ability during foreign language processing. *Journal of Cross-Cultural Psychology, 24*(4), 445-462. https://doi.org/10.1177/0022022193244005

Taube, K. T. (1995, April 18-22). *Critical thinking ability and disposition as factors of performance on a written critical thinking test*. [Paper presentation] Annual Meeting of the American Educational Research Association, San Francisco, CA.

Tian, J., & Low, G. (2011). Critical thinking and Chinese university students: A review of the evidence. *Language, Culture and Curriculum, 24*(1), 61–76. https://doi.org/10.1080/07908318.2010.546400

Tsuruta, Y. (2013). The knowledge society and the internationalization of Japanese higher education. *Asia Pacific Journal of Education, 33*(2), 140–155. http://dx.doi.org/10.1080/02188791.2013.780674

Appendix A

Sample acceptable critical responses

	Learning English	Endangered Animals
Seeking clarity	・ How much "learning" is enough? ・ "Success" has different meanings to different people.	・ What is the definition of "saved"? ・ Why all?
Challenging the logic	・ Just learning English will not lead to success. ・ If you are already successful in your own career, you don't need to learn English.	・ It's impossible to save and take care of all endangered animals. ・ If there's only one left of an animal, we can't do anything.
Presenting an alternative viewpoint	・ People who can speak English can help those who can't. ・ Getting knowledge is more important.	・ New animals can replace them. ・ To save children in developing countries is more important than saving endangered animals.
Challenging an assumption	・ Lots of people do not speak English but are successful/rich. ・ Not all jobs require English.	・ Not everyone cares about animals. ・ Animals probably aren't happy under our protection.

Appendix B

Instructions given to students

Model presented in both Practice A and B:

S16-

Critical Thinking Example

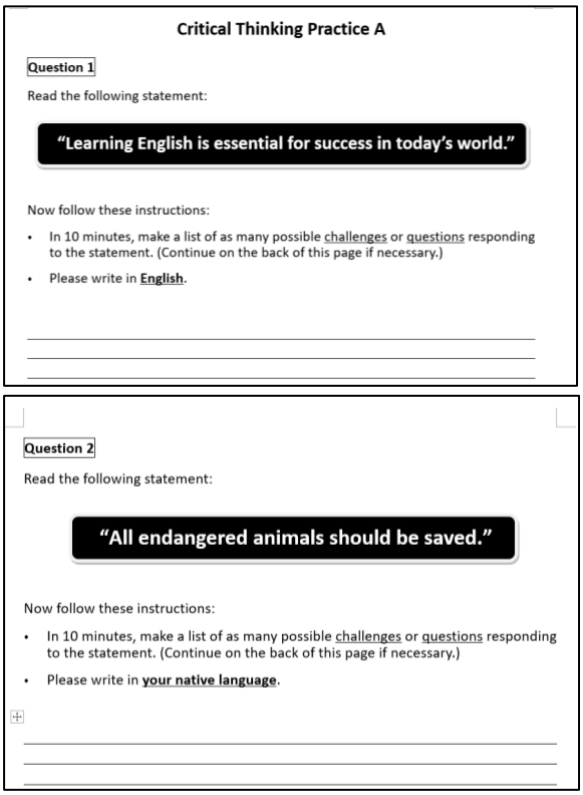
Read the following statement:

"Dogs make the best pets since they are loyal and friendly."

Below is a list of challenges or questions in response to this statement:

- *What do you mean by "loyal?"*
- *Cats can clean themselves, so they don't usually smell. However, you have to give a dog a bath, which is time consuming and can be messy.*
- *You need to walk a dog twice a day. If you're sick or don't like going outside, that can be a problem.*
- *Dogs bark loudly. This will likely disturb your neighbors.*
- *Hairy dogs shed hair. You might develop an allergy from that.*
- *Dogs need a lot of space to run around. If you have a small apartment, your dog may not be happy.*
- *Fish are better than dogs since they require less space.*
- *Some dogs are dangerous and will attack or even kill people.*
- *Could you explain "friendly" more?*
- *Dog food can be expensive. Fish food is much cheaper than dog food.*
- *If the dog is sick, you have to take it to the hospital, so medical expenses for a dog can be expensive.*
- *It'll be more difficult to find an apartment because many apartment owners won't rent to dog owners.*
- *Dogs may not want to be kept as a pet. They might feel lonely when you're not around. Also, most dogs can't pee or take a dump whenever they want because they have to wait for you to take it out for a walk.*
- *Is it really good to keep a pet, like a dog? If you have really strong feelings for your pet, you might suffer mentally when it dies.*
- *Is there any evidence dogs are more loyal than other pets, like a monkey or a cat?*
- *How do we know dogs are really loyal and friendly? Maybe they just want food.*

Practice A instructions:



Critical Thinking Practice A

Question 1

Read the following statement:

"Learning English is essential for success in today's world."

Now follow these instructions:

- In 10 minutes, make a list of as many possible challenges or questions responding to the statement. (Continue on the back of this page if necessary.)
- Please write in **English**.

Question 2

Read the following statement:

"All endangered animals should be saved."

Now follow these instructions:

- In 10 minutes, make a list of as many possible challenges or questions responding to the statement. (Continue on the back of this page if necessary.)
- Please write in **your native language**.

*Statements reversed in Practice B.

Investigating cross-linguistic similarity ratings: A Rasch analysis

David Allen[1] and Trevor Holster[2]
allen.david@ocha.ac.jp
1. Ochanomizu University, Tokyo
2. Fukuoka University, Fukuoka
https://doi.org/10.37546/JALTSIG.TEVAL25.1-3

Abstract

A robust finding in psycholinguistics is that cognates and loanwords, which are words that typically share some degree of form and meaning across languages, provide the second language learner with benefits in language use when compared to words that do not share form and meaning across languages. This *cognate effect* has been shown to exist for Japanese learners of English; that is, words such as *table* are processed faster and more accurately in English because they have a loanword equivalent in Japanese (i.e., テーブル /te:buru/ 'table'). Previous studies have also shown that the degree of phonological and semantic similarity, as measured on a numerical scale from 'completely different' to 'identical', also influences processing. However, there has been relatively little appraisal of such cross-linguistic similarity ratings themselves. Therefore, the present study investigated the structure of the similarity ratings using Rasch analysis, which is an analytic approach frequently used in the design and validation of language assessments. The findings showed that a 4-point scale may be optimal for phonological similarity ratings of cognates and a 2-point scale may be most appropriate for semantic similarity ratings. Furthermore, this study reveals that while a few raters and items misfitted the Rasch model, there was substantial agreement in ratings, especially for semantic similarity. The results validate the ratings for use in research and demonstrate the utility of Rasch analysis in the design and validation of research instruments in psychology.

Keywords: Rasch, cross-linguistic similarity, loanwords, ratings, Japanese, English

It is well known that words that share meaning and form across languages typically offer an advantage in language learning and use (e.g., *coffee, koffie,* コーヒー, *Kaffee,* and *café*). This is commonly referred to in psycholinguistics as *the cognate effect*. Linguistically speaking, such words include cognates and loanwords, which derive from the same source in etymologically related and unrelated languages, respectively. However, while this distinction between loanwords and cognates is important for linguists, it is of little consequence for language learners. Regardless of whether a word is a cognate or a loanword, if it shares meaning and form across languages, it is likely to provide a benefit in processing relative to words that do not share form and meaning. Hence, in psycholinguistics such words are generally referred to as cognates, and the effect is known as the cognate effect.

In Japanese, there are thousands of loanwords that derive from English or which share sound and meaning with English words (e.g., *coffee* and コーヒー /ko:hɪ:/ 'coffee'). In fact, half of the most common words in English have been borrowed into Japanese, and a quarter of the most common words have a commonly known loanword in Japanese (Allen, 2019c). Studies have shown that English words that have Japanese loanword equivalents that share some degree of form and meaning across the languages (i.e., Japanese-English cognates) are processed faster and more accurately than words that do not (i.e., noncognates). Specifically, studies have shown that Japanese learners of English read cognates faster when presented in isolation (e.g., Miwa et al., 2014) or in sentence context (Allen et al., 2021), and they produce cognates faster when naming pictures in English (e.g., Hoshino & Kroll, 2008). Studies have also demonstrated this effect in tests of receptive lexical knowledge (e.g., Allen, 2019a, 2019b).

The Bilingual Interactive Activation Plus (BIA+) model provides the most widely accepted explanation for how the cognate effect arises in language use (Dijkstra & Van Heuven, 2002), though the Multilink model builds on the BIA+ and extends the explanation from word recognition to word production and translation (Dijkstra et al., 2019). These models assume that during language use, word elements related to orthography, phonology and semantics become activated, the combination of which leads to word recognition and production. For instance, when an English speaker sees the word *hat*, the orthographic units h-a-t become activated, followed by the phonemes associated with them /h/, /a/, and /t/, which in turn activate the lexical representation *hat* along with the semantic representation of 'hat' that is associated with it. When a Japanese speaker of English reads the word *hat*, the same process occurs, but linguistic components in Japanese that overlap in phonology and semantic features also become activated in parallel. If a word exists in the lexicon that is similar in form and meaning to the English word (i.e., ハット /haʔto/ 'hat'), the resulting activation of shared phonological and semantic features is believed to underlie the boost in processing of the English word. In short, when words with similar phonological and semantic features exist in the lexicon, they co-activate one another, which typically leads to a benefit in processing, though the effect will vary according to the task.

Measuring cross-linguistic similarity

A key feature of the processing of cognates is that the extent of the processing advantage appears to vary according to the extent of cross-linguistic similarity. That is, rather than an 'all-or-nothing' cognate effect, it is really a 'gradient' cognate effect which is further defined by the extent of cross-linguistic formal (orthographic and/or phonological) and semantic similarity (Allen et al., 2021; Dijkstra et al., 1999, 2010). In terms of orthography, words in same-script languages, such as Dutch and English, can either share identical (e.g., *bed-bed*) or similar (e.g., *apple-appel*) orthographic form. The difference between words' orthographic forms can be computed objectively using a formula such as Van Orden's (1987) orthographic similarity measure or Normalized Levenshtein Distance (e.g., see Dijkstra et al., 2019; Schepens et al., 2012). The degree of overlap as measured by these formulae has been shown to predict response times in word recognition tasks where participants read words in their second language (e.g., Dijkstra et al., 2019; Van Assche et al., 2009). Furthermore, subjective ratings of orthography have also been shown to predict word recognition times in such tasks. These involve participants rating the similarity of a translation pair using a scale of similarity, for instance using a 7-point Likert-type scale ranging from 'no similarity' to 'perfect similarity' (e.g., Dijkstra et al., 2010).

In different-script languages, such as Japanese and English, orthographic overlap is essentially set at zero, leaving phonological overlap as the sole formal feature for noticing cross-linguistic similarity. Measuring phonological overlap, however, is much less straightforward than measuring orthographic overlap. This is because the English and Japanese sound systems are very different: Two words (e.g., *hat* and ハット/haʔto/ 'hat') may share some similar phonological features, such as /h/, /a/, and /t/, though these are not pronounced identically across languages. Moreover, English is stress-accented while Japanese is pitch-accented, which creates additional differences for loanwords and their translations. Although there have been attempts to create an objective measure of phonological similarity (e.g., Miwa et al., 2014), due to the inherent difficulty in creating a precise measure, researchers have tended to use phonological similarity ratings (e.g., Allen & Conklin, 2013; Allen et al., 2021; Dijkstra et al., 2010; Miwa et al., 2014). Thus, as described above for orthographic overlap, bilinguals rate the similarity of translation pairs using a Likert-type scale and the average rating for each pair is used as a measure of the two words' phonological overlap. Studies using a subjective measure of phonological similarity have typically found that these ratings significantly predict word recognition times, such that English words with more similar sounding Japanese loanwords are recognized faster than those with less similar sounding loanword equivalents (e.g., Allen & Conklin, 2013; Allen et al., 2021; Miwa et al., 2014).

In addition to formal similarity, researchers must also consider translation equivalence or the semantic similarity of translations. During the process of word recognition, the semantic features associated with words become activated. Words with greater overlap across languages are expected to co-activate to a greater extent due to the greater activation of shared conceptual features. Consequently, measures of cross-linguistic conceptual equivalence, translation equivalence, and semantic similarity, have been used in studies of bilingual language processing (e.g., Allen & Conklin, 2013; Dijkstra et al., 2010; Miwa et al., 2014; Tokowicz et al., 2002). In a norming study, Tokowicz et al. (2002) found that a subjective measure of semantic similarity correlated significantly with the number of translations that the word has, context availability (i.e., how easy it is to think of a context for a word), and concreteness. That is, words that are rated as more semantically similar tend to have fewer translations, be more concrete and have more identifiable contexts of use (Tokowicz et al., 2002). Similarly, for Japanese-English translation equivalents, Allen and Conklin (2014) showed that semantic similarity correlates with the number of senses, number of translations, and concreteness of words.

Although studies have used subjective measures of orthographic, phonological, and semantic overlap, there has been little discussion as to the creation of these measurements. Many studies have used a Likert-type scale with numbers (i.e., numerical scales) and labels at the extremes of the scale (e.g., Allen & Conklin, 2013; Dijkstra et al., 2010; Miwa et al., 2014; Tokowicz et al., 2002). The wording of the extremes has varied slightly (e.g., 'completely similar' or 'exactly the same'), though this is unlikely to influence the outcomes. Moreover, most studies have used 7-point scales though some have used 5-point scales. The rationale for using a 5-point scale (Allen & Conklin, 2013, 2014) was that during piloting of the scale, raters appeared to have difficulty utilizing certain parts of the scale (i.e., points between the extremes and the middle of the scale: 2, 3, 5, 6). In other words, while responses were reasonably well distributed, it appeared that some parts of the scale were being used more than others. In other studies, formal similarity ratings were reported to be more-or-less evenly distributed over the whole scale (Dijkstra et al., 2010; Tokowicz et al., 2002).

In contrast to formal similarity ratings, studies investigating the semantic similarity of translation equivalents have found that, perhaps unsurprisingly, ratings clump together at the 'high similarity' end of the scale (Allen & Conklin, 2013; Miwa et al., 2014; Tokowicz et al., 2002). That is, although there was variation in the degree of semantic similarity of translation pairs, they were most often rated as being almost identical. In an attempt to deal with this issue, Miwa et al. (2014) collapsed the data collected about translation equivalence between English-Japanese words from a 7-point scale to a 2-point scale, that is, 'identical' (items receiving a '1' on the scale from three out of four raters, $N = 151$) and 'non-identical' (all remaining

items, $N = 99$).[1] Although there may be little direct impact on any subsequent analyses, the fact that raters tend to use some parts of a scale much more than others raises the question of whether cross-linguistic similarity is best measured using Likert-type scales or some other method (e.g., dichotomous choice), and if Likert-type scales are appropriate, how many points are optimum for measurement.

In all of the above studies, researchers have determined which method of measuring cross-linguistic similarity appears to be the best in their context, that is, for use with specific languages, items, and participants. Therefore, it is unsurprising that there is some variation in the exact method of measuring cross-linguistic similarity of translation pairs. Nevertheless, it would be prudent to further investigate the structure of cross-linguistic similarity ratings in order to better understand them and better guide future studies. To this end, the present study performs a Rasch analysis to investigate structure of ratings.

The Rasch model and objective measurement

The dichotomous Rasch model was introduced by Georg Rasch (1960) and further developed by Wright and Stone (1979). The Rasch definition of measurement requires an equal-interval scale on Stevens' (1946) hierarchy, which is achieved by conversion of raw scores to log-odds units, or logits. In this model, unidimensionality, local item independence, and data-model fit are requirements of measurement (Aryadoust, Ng, & Sayama, 2021), so empirical demonstration that these requirements have been met is a prerequisite to any Rasch based validity argument. Moreover, Rasch model measurement invariance depends on meeting the requirements of *specific objectivity* (Engelhard, 2013), where item difficulty is invariant between different samples of persons and person ability is invariant between different samples of test items. Thus, the Rasch model provides *objective measurement,* despite the inherent subjectivity of human responses.

A crucial difference between the Rasch model and other item response theory (IRT) models is that Rasch analysis functions as a confirmatory analysis of whether the dataset fits a prescriptive measurement model, whereas IRT analysis aims to fit a model to the observed data (DeMars, 2010). The standard Rasch analysis of data-model fit is through the mean-square fit statistic, provided as an information weighted *infit* and unweighted *outfit* statistic. The expected value of the mean-square statistic is 1.00, with Linacre (2009) suggesting mean-square values below 1.50 as productive for measurement, with values exceeding 2.00 unproductive. High mean-square values are known as *misfit* or *underfit*, indicating idiosyncratic responses. Excessive misfit precludes objective measurement.

Andrich (1978) and Masters (1982) introduced polytomous Rasch models, allowing analysis of whether respondents interpret rating scale categories consistently. This was extended by Linacre's (1994) many-faceted Rasch measurement (MFRM), which allows additional measurement *facets*, such as *raters*, to be analyzed along with the familiar two facets of *participants* and *items* from traditional tests. In addition to rater leniency or severity, in which different raters may systematically assign higher or lower scores for the same performance, fit statistics can also be used to diagnose idiosyncratic rating behavior, evidence of raters interpreting the rating rubric differently. In a seminal study of language performance assessments, McNamara (1996) conducted a fit analysis of raters which revealed that they often behave idiosyncratically. Rasch analysis has since become an essential component in the implementation of high-stakes language assessments, where raters' scores can be automatically adjusted according to rater severity to improve fairness in terms of scoring validity.

In other research designs, misfitting raters (and items) may be identified using Rasch analysis and thereafter retained or removed. However, while it is tempting to exclude idiosyncratic responses from analyses in order to improve data-model fit, a process Davidson (2000) criticized as "statistical determinism", many constructs of interest to linguists cannot be disentangled from subjective human judgements, so some level of idiosyncratic behavior from human raters needs to be accepted. Furthermore, misfitting responses may in fact provide important insights into the nature of the construct being investigated, not a nuisance to be removed. For example, if the rater pool has a majority of raters with exposure to a particular language variety, raters from other backgrounds may misfit relative to the dominant language variety. This can be identified by very low mean-square values, known as *overfit,* which show that the raters generally display very high agreement, indicating redundancy in the data. Consequently, because the mean-square statistic is constrained to have an average value close to 1.00, highly consistent raters will cause other raters to appear to be relatively inconsistent. In this way, Rasch analysis may provide additional insight into the construct during instrument development. Moreover, despite the inherent human subjectivity in participant responses, Rasch derived logit measures provide an objective measurement scale if Rasch data-model fit is adequate.

Research questions

To investigate the structure of previously collected cross-linguistic similarity ratings, a Rasch analysis was performed and guided by the following three research questions. The three questions were investigated firstly for phonological similarity ratings and then for semantic similarity ratings.

1. What is the optimum number of points on the scale when investigating cross-linguistic similarity?
2. Are there any items or participants that display significant unexpected variation in responses?
3. Are raw scores a sufficient approximation of an interval scale to provide useful measures of loanword similarity?

Method

The data in this study is taken from Allen et al. (2021). Twenty-nine female undergraduates at a Japanese university took part in the rating study. Participants had a mean score of 54 on the *Vocabulary Size Test* (Nation & Beglar, 2007; SD = 9.7) suggesting an estimated vocabulary size of 5400 words, which is indicative of intermediate English reading ability. They completed two rating tasks for 108 English and Japanese loanword translations (e.g., advice – アドバイス). First, participants rated the phonological similarity of the words on a 7-point numerical scale from 1 ("Completely different") to 7 ("Identical"). Next, they rated the semantic similarity of the same words using the same method. Prior to rating, participants were given a number of brief examples indicating how some word pairs could be perceived as having relatively high or low phonological overlap (e.g., *tennis*-テニス / tenɪsu/ and *radio*-ラジオ/radʒɪo/, respectively) and relatively high or low semantic overlap (e.g., *radio*-ラジオ/ radʒɪo / 'radio' and *side*-サイド/saɪdo/ 'side (dish)', respectively).

Importantly, items in this particular rating study included cognates (i.e., English words paired with *katakana* translation equivalents; e.g., *tennis*-テニス) but not noncognates (i.e., English words paired with *kanji* translation equivalents; e.g., *clock*-時計 /tokei/ 'clock'). Therefore, ratings at the extremely dissimilar end of the scales, which would indicate word pairs that are completely different in either form or meaning, were not expected. Also, although such rating tasks are typically completed using online survey software, logistical issues at the time meant that the ratings were collected using a pencil-and-paper method. The main difference resulting from the use of this method was that rather than being presented randomly for all participants, word pairs were presented in alphabetical order.

Data was analyzed using Winsteps version 4.6.2 (Linacre, 2020) using the Andrich rating scale model and dichotomous Rasch model.

Results

Phonological similarity ratings analysis

Optimal number of scale categories

An analysis of the original 7-point scale was followed by analysis of collapsed scales, revealing that a 4-point scale may be optimal. Figure 1 shows the category probability curves for the original 7-point scale and a 4-point scale with the lower four categories collapsed. The curves show the range across which each category is most likely to be observed, with the rating categories of "2", "3", and "4" covering a small range in the original 7-point scale. Table 1, which reports the category structure of the two scales, shows that ratings of "1" to "4" only accounted for 21% of responses, and thus provide relatively little information compared to the higher categories. The lower response categories can also be seen to be misfitting, with mean-square infit and outfit values exceeding 1.00 for the 7-point scale. When these categories were collapsed into a single category, the mean-square fit statistics generally improved, with maximum infit and outfit values of 1.18 and 1.26, respectively, compared with 2.30 and 3.27 for the 7-point scale. This suggests that seven rating categories are too many and that a 4-point scale will result in more consistent rating behavior for rating phonological similarity of cognate words.

Figure 1

Category probability curves for 7-point rating scale (left) versus 4-point scale (right) of phonological similarity

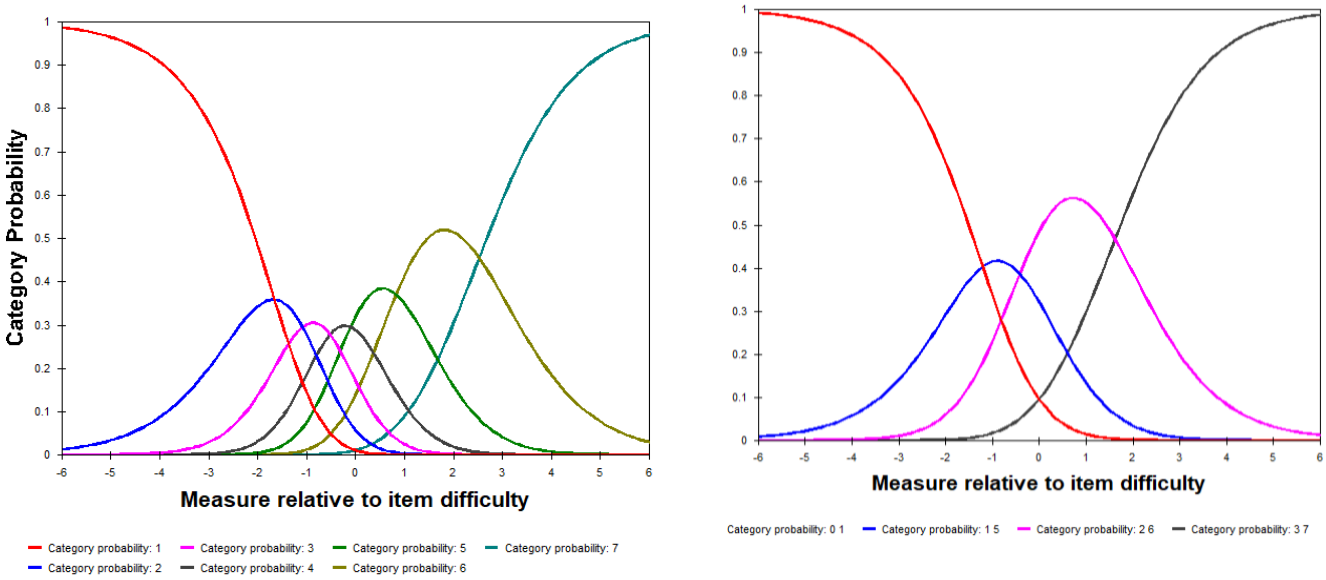


Table 1

Summary of scale category structure for phonological similarity

Rating	Score		Count		% Observed		*M* (Logits)		Infit *MS*		Outfit *MS*		Andrich Threshold		Category Measure	
	7-Pt	4-Pt	7-Pt	4-Pt	7-Pt	4-Pt	7-Pt	4-Pt	7-Pt	4-Pt	7-Pt	4-Pt	7-Pt	4-Pt	7-Pt	4-Pt
1	1	0	33		1		-0.08		2.30		3.27		n.a.		(-3.06)	
2	2	0	87		3		-0.33*		1.22		1.25		-1.66		-1.68	
3	3	0	190		6		-0.06		1.07		1.20		-1.08		-0.87	
4	4	0	361	671	12	21	0.25	-1.56	0.96	1.18	0.96	1.26	-0.52	n.a.	-0.22	(-2.55)
5	5	1	736	736	23	23	0.80	- 0.63	0.98	0.98	0.93	1.01	-0.10	-1.21	0.56	-0.89
6	6	2	1094	1094	35	35	1.55	0.44	0.84	0.91	0.83	1.06	0.83	-0.41	1.82	0.73
7	7	3	631	631	20	20	2.51	1.71	0.97	0.87	0.96	1.24	2.53	1.62	(-3.75)	(-2.81)

Note. * indicates disordered category

Dimensionality, dependency, and data-model fit

Unidimensionality, local item independence, and acceptable data-model fit are requirements for the Rasch measurement model. Considering unidimensionality, Reckase (1979) stipulated 20% variance explained by the major dimension as the minimum requirement, while Linacre (2016) emphasized that the relative size of the Rasch dimension compared to sub-dimensions is a major consideration. Dimensionality is typically investigated through principal components analysis of residuals (PCAR), which differs from the normal procedure of principal components analysis (PCA) in that the expected response is subtracted from the observed response before conducting the analysis, allowing comparison of the size of any sub-dimensions relative to the size of the Rasch dimension. PCAR found the Rasch dimension to account for 49.90% and 55.70% of variance for the 7-point and 4-point scales, respectively. The largest contrasting dimensions, representing 8.20% and 5.10% of variance, respectively, were approximately 16% and 9% of the Rasch dimensions, sufficiently small to justify analysis as a unidimensional instrument (Linacre, 2016).

Item dependency was investigated through analysis of dependent item pairs based on standardized item residual correlations, with values greater than .70 raising serious concern (Linacre, 2020). Nine item pairs had correlations exceeding .70 for both the 7-point scale and the collapsed 4-point scale (Table 2), while eight pairs exceeded this value for the 7-point scale alone and five pairs for the 4-point scale alone.

A small proportion of these dependent pairs displayed phonological similarities, for example “wind” and “wing”, which may explain their high correspondence. Additionally, these dependent items were typically adjacent in the alphabetical list,

which may have further highlighted phonological similarities to raters. However, the majority of items do not share notable phonological similarities, so it is unclear why they exhibit high item dependency.

Table 2

Most dependent item correlations

	7-point	4-point	Item 1	Item 2
Co-occurring	0.87	0.80	5 banana	42 idea
	0.82	0.78	93 sugar	94 summer
	0.71	0.82	85 shoe	88 sock
	0.80	0.73	27 drama	41 hotel
	0.75	0.78	77 rail	81 saddle
	0.76	0.66	24 desk	39 head
	0.76	0.73	106 wind	107 wing
	0.74	0.73	12 case	13 chain
	0.73	0.71	1 advice	5 banana

The effect of item dependency on logit measures was investigated by removing all the items occurring in any of the 20 most dependent pairs in either analysis and determining logit difficulties of the remaining items to use as anchoring values. Because some items occurred in multiple dependent pairs, 24 dependent items were removed, leaving 84 anchoring items. The 24 dependent items were then returned to the analysis and item difficulties compared between the anchored and unanchored analyses (see Linacre, 2020). The maximum absolute difference in item difficulty was 0.12 logits, with a mean absolute value of 0.03 logits and a Pearson correlation of approximately 1.00. Item dependency thus did not have a substantively or statistically significant effect on measurement invariance.

Data-model fit was investigated by analysis of summary statistics for persons and items, reported in the Infit *MS* and Outfit *MS* columns of Table 3 and Table 4. Both persons and items showed misfit for both the 7-point scale and 4-point scale. Figure 2 shows the empirical and modeled item characteristic curves (ICCs). The 7-point scale sharply diverges from the modeled curve below the rating category of 4, while the empirical curve for the 4-point scale much more closely follows the modeled curve because the most misfitting responses have been restricted to a single category. Figure 3 shows the item pathway maps for mean-square outfit and infit, using the 4-point scale. The vertical axis shows logit measures. Words with lower raw scores are higher on the map, indicating greater perceived phonological difference. Both panels show a trend of more similar words tending to overfit, with less similar words misfitting. The overfitting items exaggerate the relative misfit of the more difficult (i.e., less similar) items, which are sensitive to a very small number of idiosyncratic responses.

Table 3

Person summary statistics for phonological ratings (N = 29)

	7-Point Scale						4-Point Scale					
			Infit		Outfit				Infit		Outfit	
	Logit	*SE*	*MS*	ZSTD	*MS*	ZSTD	Logit	*SE*	*MS*	ZSTD	*MS*	ZSTD
M	1.25	0.11	1.12	-0.51	1.06	-0.75	0.02	0.14	1.06	-0.46	1.15	-0.30
SD	1.04	0.03	0.96	4.51	0.87	4.54	1.35	0.02	0.59	4.13	0.96	4.19
Max.	3.23	0.17	5.13	9.91	4.44	9.90	2.32	0.18	2.77	8.86	5.39	9.91
Min.	-0.75	0.08	0.26	-7.43	0.25	-7.66	-2.47	0.12	0.27	-8.88	0.29	-8.25
Reliability:		7-Point Scale .98			4-Point Scale .99							
Separation:		7-Point Scale 7.40			4-Point Scale 8.12							

Table 4

Item summary statistics for phonological ratings (N = 108)

	7-Point Scale						4-Point Scale					
			Infit		Outfit				Infit		Outfit	
	Logit	*SE*	*MS*	ZSTD	*MS*	ZSTD	Logit	*SE*	*MS*	ZSTD	*MS*	ZSTD
M	0.00	0.21	1.02	-0.04	1.06	0.06	0.00	0.27	1.00	-0.10	1.15	0.10
SD	0.53	0.02	0.47	1.60	0.54	1.79	0.74	0.01	0.38	1.39	0.83	1.75
Max.	1.48	0.26	2.23	3.54	2.73	4.72	2.25	0.32	2.47	3.70	6.26	6.81
Min.	-1.11	0.17	0.25	-3.70	0.33	-3.24	-1.35	0.26	0.44	-2.77	0.47	-2.51
Reliability:		7-Point Scale .82			4-Point Scale .85							
Separation:		7-Point Scale 2.14			4-Point Scale 2.34							

Figure 2

Empirical and modeled ICC for 7-category scale (left) versus 4-point scale (right) of phonological similarity

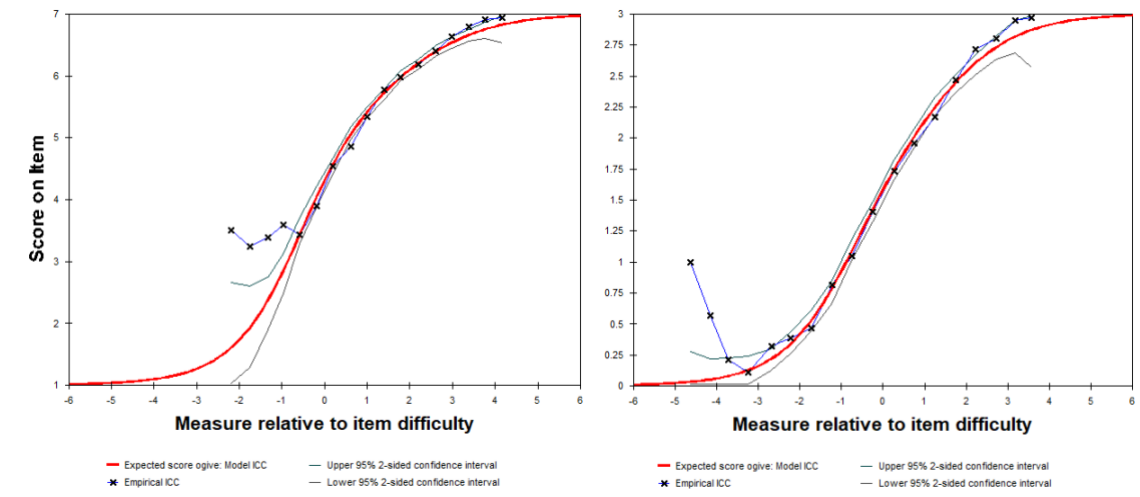


Figure 3

Item pathway maps of outfit (left) and infit (right) for the collapsed 4-point scale

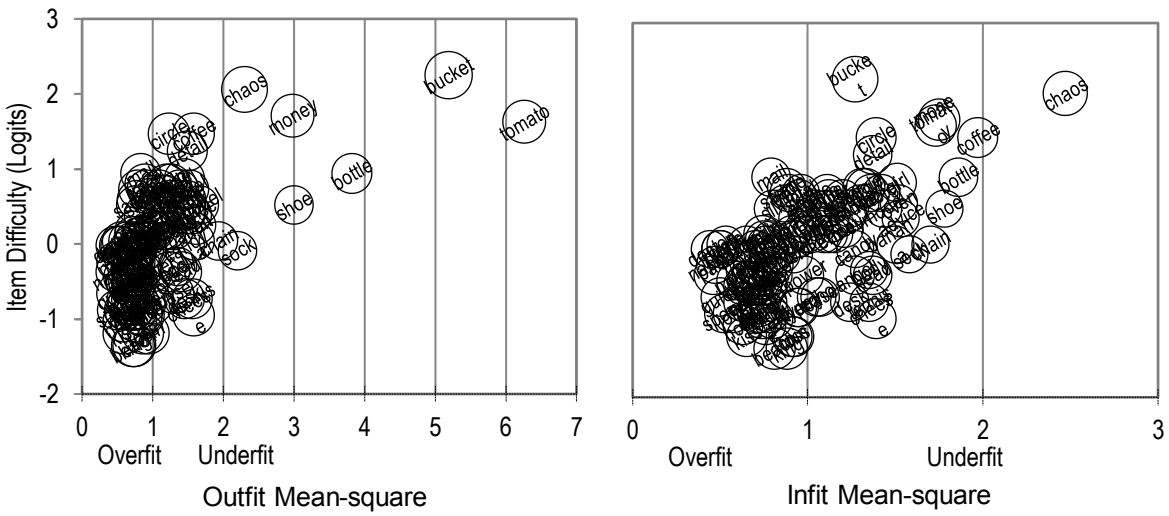
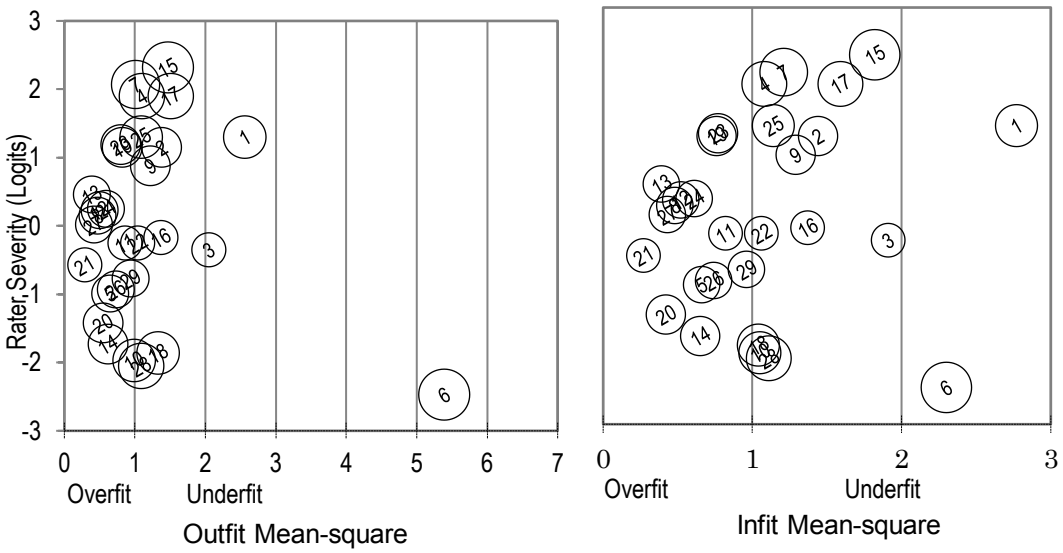


Figure 4 shows the person pathway maps, with Raters 1, 3, and 6 having outfit mean-square values exceeding 2.00 and high infit values. Raters 17 and 15 also showed high infit values, but these raters were extremely lenient, meaning they judged

nearly all words to be highly similar. Rater 6 was extremely severe, rating words as much less similar than the other raters. It is possible that this rater has some background characteristics that would explain this difference.

Figure 4

Person pathway maps of outfit (left) and infit (right) for the collapsed 4-point scale

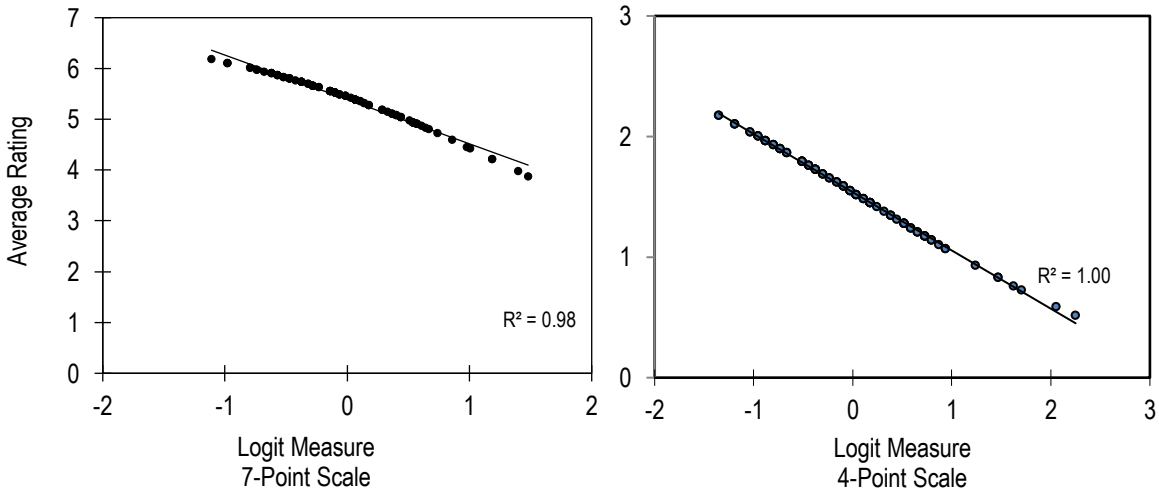


To further investigate rater idiosyncrasy, an analysis was conducted of the most unexpected responses, that is, those responses with the largest standardized residuals from the rescored scale. Of the 20 most unexpected responses, nine came from Rater 6, further highlighting her as behaving extremely unusually relative to the group. Although removing idiosyncratic raters is not advised (Davidson, 2000), a reanalysis was performed by removing Raters 1, 3, and 6. This revealed that while the fit statistics improved, especially for items, the dataset was still noisy. Overall, many highly similar words continued to overfit, which exaggerated the relative disagreements over the misfitting words.

Figure 5 compares item logit measures with mean ratings for each item on the rating scale, with a nearly perfectly linear relationship for the 7-point scale and the 4-point scale data. For this particular dataset, the raw scores and logit measures are effectively interchangeable. This linear relationship occurred because no items approached the extremes of either rating scale, in which case the relationship would inevitably become increasingly non-linear.

Figure 5

Comparison of item logit measures and mean ratings for the 7-point (left) and 4-point (right) scales



Semantic similarity ratings analysis

Optimal number of scale categories

An analysis of the original 7-point scale was followed by analysis of collapsed scales, revealing that a dichotomous (2-point) scale may be optimum. Figure 6 shows the category probability curves for the original 7-point scale ratings and for the dichotomous scale ratings, in which the lower six categories were collapsed. The categories for the 7-point scale are not well defined, ratings of "2", in particular, are never the most probable response. Table 5 illustrates that this is due to categories below "5" being used extremely rarely, constituting only 10% of the responses. It is also clear from the Infit *MS* and Outfit *MS* columns in Table 5 that the lower categories exhibited worrying levels of misfit, but that the category of "6" was highly overfitting, with an outfit mean-square value of 0.52. Collapsing the data into dichotomous ratings resulted in generally improved data-model fit.

Figure 6

Category probability curves for 7-point rating scale (left) versus 2-point scale (right) of semantic similarity

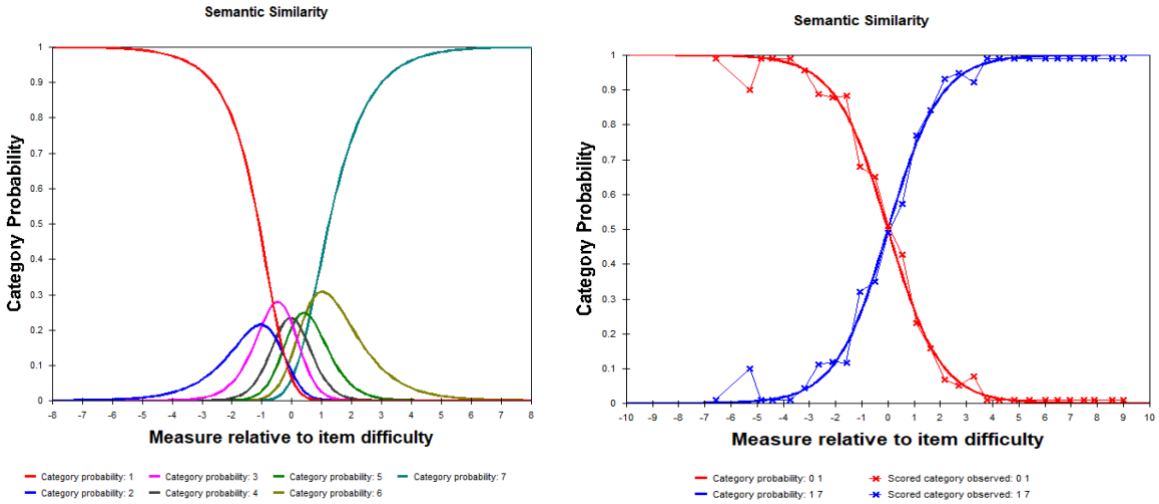


Table 5

Summary of scale category structure for semantic similarity

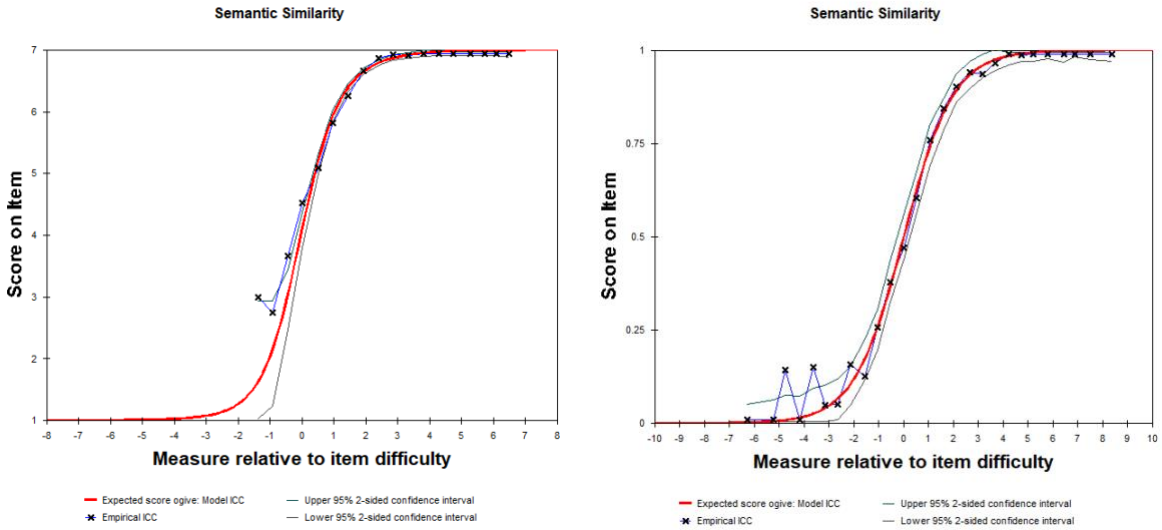
Rating	Score		Count		% Observed		*M* (Logits)		Infit *MS*		Outfit *MS*		Andrich Threshold		Category Measure	
	7-Pt	2-Pt	7-Pt	2-Pt	7-Pt	2-Pt	7-Pt	2-Pt	7-Pt	2-Pt	7-Pt	2-Pt	7-Pt	2-Pt	7-Pt	2-Pt
			23		1											n.a.
1	1	0					0.06		1.58		2.27		NONE	n.a.	(-2.05)	
2	2	0	26		1		0.28		1.45		1.76		-0.25	n.a.	-1.03	n.a.
3	3	0	83		3		0.48		1.24		1.72		-0.99	n.a.	-0.46	n.a.
4	4	0	144		5		0.79		1.16		1.76		-0.07	n.a.	-0.04	n.a.
5	5	0	294		9		0.93		0.99		0.91		0.12	n.a.	0.40	n.a.
6	6	0	639	1209	20	39	1.30	-0.83	0.93	0.98	0.52	0.93	0.47	n.a.	1.02	n.a.
7	7	1	1922	1922	61	61	2.29	1.94	0.98	1.00	0.97	1.15	0.73	n.a.	-2.23	n.a.

Note. One missing response was recorded.

Figure 7 shows the modeled and empirical ICCs, with very narrow confidence intervals for the higher categories on the scale, but large confidence intervals for categories below 5. The dichotomous ratings closely follow the modeled curve for ratings above 0.25, but the very low ratings, which indicate large semantic differences, diverge from the model. These results indicate that raters were unable to effectively distinguish seven rating categories, so a dichotomous scale seems more appropriate.

Figure 7

Empirical and modeled ICC for 7-category scale (left) versus 2-point scale (right) of semantic similarity



Dimensionality, dependency, and data-model fit

Unidimensionality was investigated through PCAR analysis, which showed that the Rasch dimension accounted for 39.70% and 42.40% of variance for the 7-point and dichotomous scales, respectively. The largest contrasting dimensions represented 7.70% and 5.70% of variance, respectively, approximately 19% and 13% of the Rasch dimensions. These values justify analysis as a unidimensional instrument (Linacre, 2016).

Item dependency was investigated through examination of the standardized residual correlations for item pairs. Eight item pairs were highly correlated in both analyses (Table 6), with 12 items highly correlated only in the 7-point scale and 10 only in the dichotomous scale, including two items having very high negative correlations in the dichotomous scale. Negative correlations were for *tomato-moment* and *tomato-noise*, revealing that while *tomato* and トマト were rated as highly similar in English and Japanese, *moment* and モーメント, and *noise* and ノイズ, were rated as very different across languages. The 35 items included in the dependent pairs (M = -0.54, SD = 0.70) were substantively and statistically significantly more similar than the independent items (M = 0.26, SD = 0.57), $t(56)$ = -5.94, $p < .001$). In the vast majority of cases, therefore, these high correlations reflect the tendency for raters to rate English and Japanese word pairs as highly similar in terms of meaning.

Table 6

Most dependent item correlations for semantic ratings

	7-point	Dichotomous	Item 1	Item 2
Co-occurring	0.85	1.00	86 silk	90 spoon
	0.80	1.00	32 fruit	46 knife
	0.93	0.77	21 coffee	38 guitar
	0.86	0.86	18 cherry	47 lion
	0.83	0.85	94 summer	97 tennis
	0.78	0.83	69 plan	70 plant
	0.80	0.80	61 monkey	93 sugar
	0.78	0.74	21 coffee	46 knife

Data-model fit was examined through summary statistics for raters (Table 7) and items (Table 8). The respective person reliability indices of .93 and .96 for the 7-point scale and dichotomous scale give separation indices of 3.72 and 5.38, indicating very high confidence that raters were statistically significantly different in severity. Both analyses found a range of rater severity exceeding 5 logits, a substantively very large difference, comparable to the range of item difficulty.

However, in this study, all raters judged all items, so the averaged raw ratings avoid this problem. Rasch logit values automatically adjust for rater severity, but this is conditional upon acceptable data-model fit. Tables 7 and 8 also show concerning levels of misfit for the 7-point scale, with respective infit and outfit values of 1.27 and 1.07 for raters and 1.11 and 1.07 for items. The dichotomous ratings are close to the expected value of 1.00, with respective infit and outfit values of 0.99 and 1.02 for both raters and items. This makes it clear that raters did not interpret the intermediate categories on the rating scale consistently.

Table 7

Person summary statistics for semantic ratings (N = 29)

	7-Point Scale						2-Point Scale					
			Infit		Outfit				Infit		Outfit	
	Logit	*SE*	*MS*	ZSTD	*MS*	ZSTD	Logit	*SE*	*MS*	ZSTD	*MS*	ZSTD
M	1.81	0.14	1.27	0.59	1.07	0.10	0.87	0.29	0.99	-0.11	1.02	-0.01
SD	0.83	0.09	0.68	2.41	0.54	2.11	1.70	0.11	0.22	1.70	0.60	1.63
Max.	4.59	0.57	3.62	6.32	2.86	7.23	4.96	0.76	1.62	3.70	3.08	4.10
Min.	0.43	0.07	0.54	-2.99	0.48	-2.61	-1.71	0.23	0.74	-2.97	0.48	-2.43
Reliability:		7-Point Scale .93			2-Point Scale .96							
Separation:		7-Point Scale 3.72			2-Point Scale 5.38							

Table 8

Item summary statistics for semantic ratings (N = 108)

	7-Point Scale						2-Point Scale					
			Infit		Outfit				Infit		Outfit	
	Logit	*SE*	*MS*	ZSTD	*MS*	ZSTD	Logit	*SE*	*MS*	ZSTD	*MS*	ZSTD
M	0.00	0.26	1.11	0.22	1.07	0.19	0.00	0.52	0.99	-0.02	1.02	0.15
SD	0.72	0.11	0.58	1.24	0.68	1.28	1.44	0.10	0.23	1.00	0.64	0.82
Max.	1.86	0.69	3.74	4.14	3.41	4.58	4.61	1.05	1.62	2.67	4.12	2.47
Min.	-1.9	0.13	0.31	-2.66	0.30	-2.12	-3.41	0.46	0.42	-3.19	0.22	-2.00
Reliability:		7-Point Scale .80			2-Point Scale .85							
Separation:		7-Point Scale 2.00			2-Point Scale 2.41							

Rater and item fit were also investigated through examining pathway maps. Figure 8 shows the pathway map for raters, with Raters 3, 15, and 6 being of most concern. Rater 3 was extremely strict, judging most words as highly dissimilar across languages, so the outfit statistics for this rater would have been sensitive to a few outlying responses. However, Raters 6 and 15 were near the middle of the range of severity, so these two raters are perhaps of more concern in terms of idiosyncratic responses. These findings for semantic ratings overlap somewhat with those for phonological ratings, where Raters 3 and 6 were both identified as behaving unusually relative to the group. Looking at item fit, Figure 9 shows the outfit and infit item pathway maps for the dichotomous ratings, with many overfitting items and two seriously misfitting items. The item misfit is largely confined to the outfit statistic, reflecting that the most misfitting word pairs, *circle*-サークル and *water*-ウオーター, were near the upper and lower extremes of the difficulty range. This likely reflects the general homogeneity within the responses, which makes these two items, which were rated less consistently, poorly fit the model.

Figure 8

Person pathway maps of outfit (left) and infit (right) for the dichotomous ratings of semantic similarity

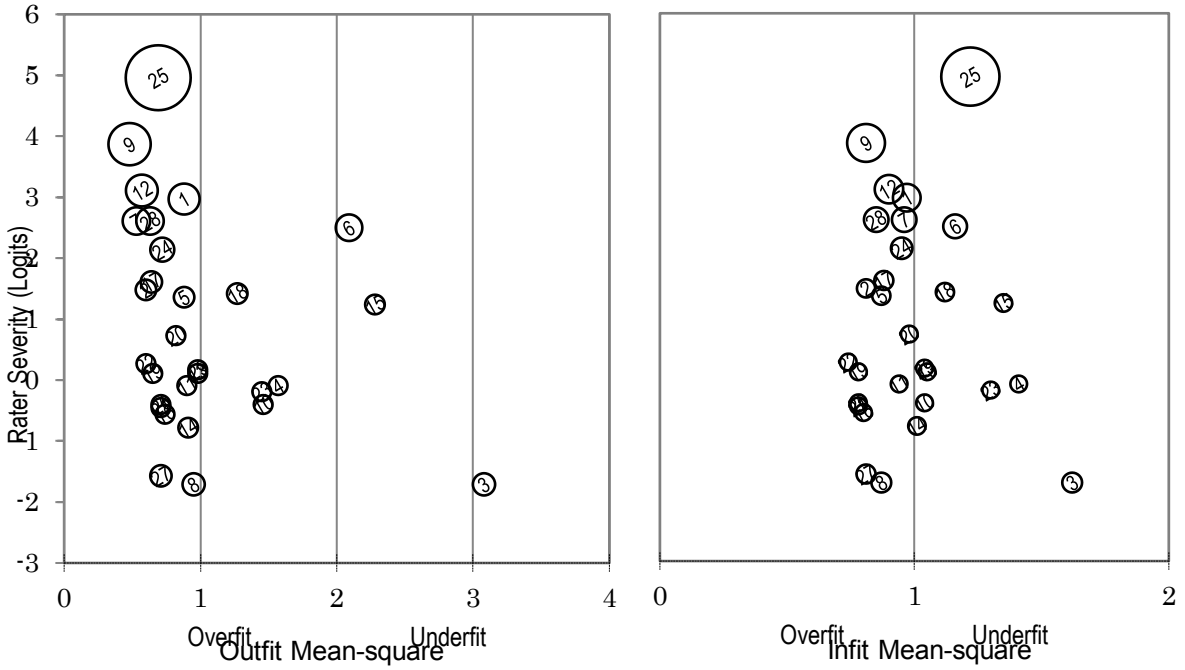


Figure 9

Item pathway maps of outfit (left) and infit (right) for the dichotomous ratings of semantic similarity

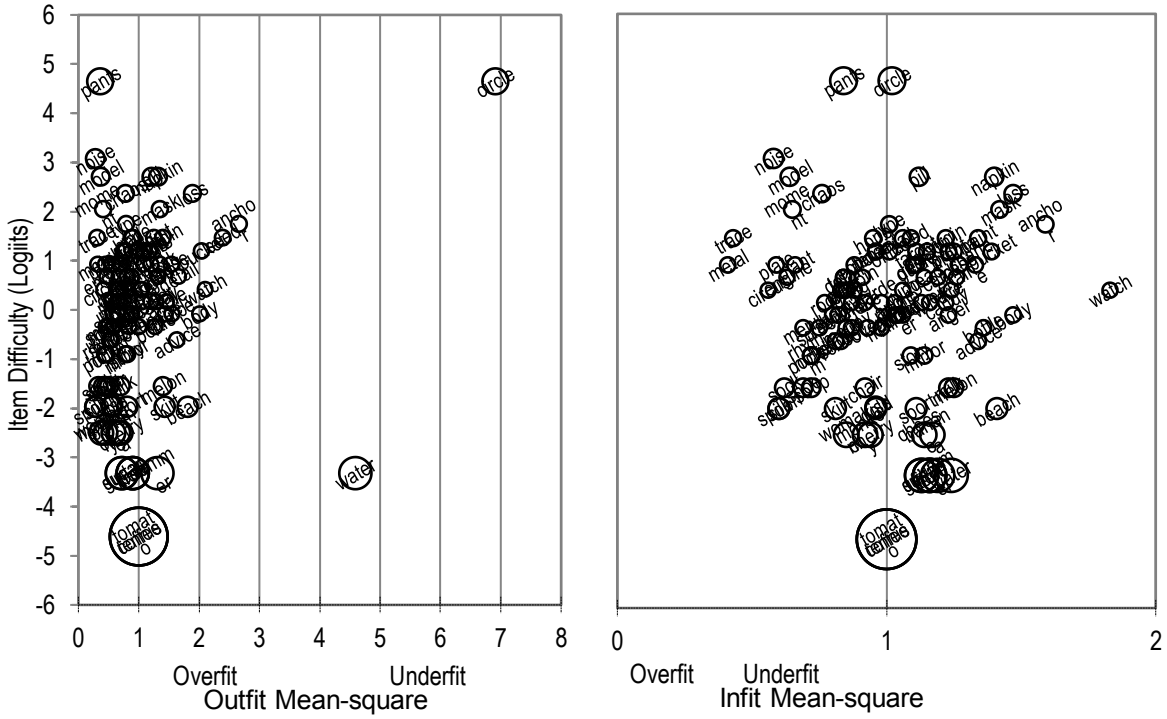
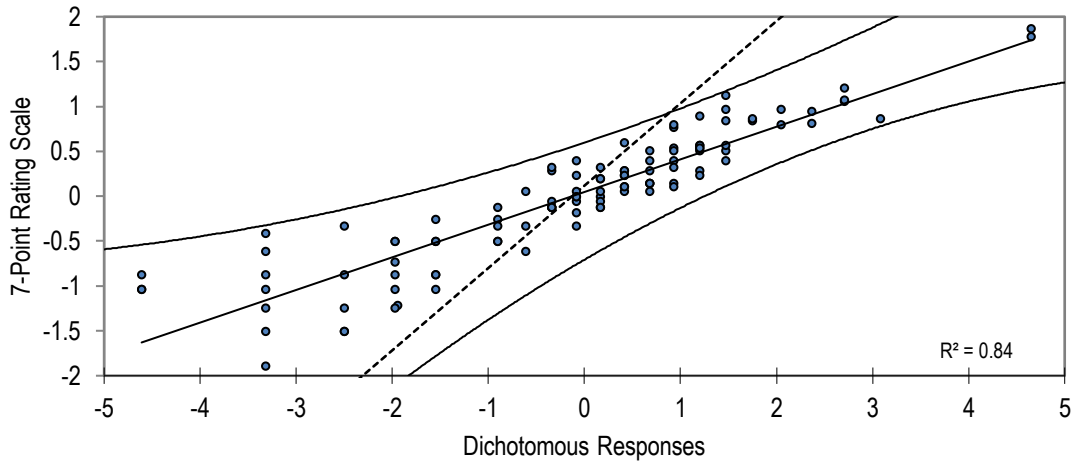


Figure 10 compares item difficulty (i.e., semantic similarity rating) from the 7-point rating scale and dichotomous ratings. No items fall outside the 95% confidence intervals, with the 84% shared variance reflecting a raw correlation of .93, rising to 1.00 after disattenuation for reliability. Figure 11 compares item raw scores with logit measures. The left-hand panel, showing the 7-point scale, illustrates that rank-ordering is retained between raw scores and Rasch generated logit measures,

but that the relationship became increasingly distorted as the maximum score of 7 was approached. This distortion was not observed for the phonological data because extreme scores were not observed. The right-hand panel compares logit values from the dichotomous rescaling with raw scores from the original 7-point scale, revealing shared variance of 74% for the raw ratings and logits from the rescaled dichotomous ratings. Although such extreme rescaling typically results in a considerable reduction in shared variance, this was not observed because ratings below "5" were rarely observed.

Figure 10

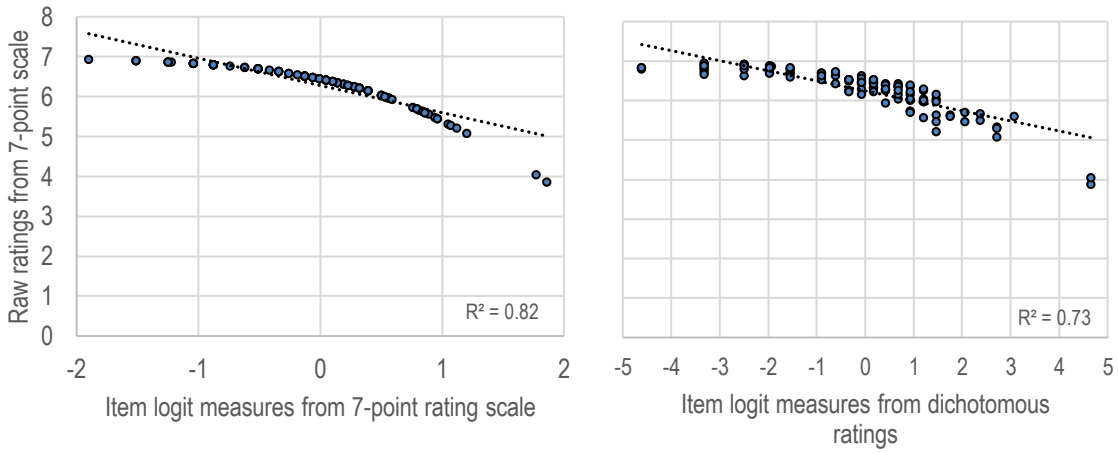
Comparison of semantic similarity ratings from 7-point rating scale and rescaled dichotomous data



Note. The upper and lower solid lines show the 95% confidence intervals, with the linear trendline shown in solid. The differing slopes of the empirical trendline and the dashed identity line show that the logit scale has been stretched by the rescaling of the responses.

Figure 11

Comparison of 7-point rating scale (left) and dichotomous (right) item scores with logit measures



Discussion

The present study illustrates the potential of Rasch analysis to evaluate the reliability of an instrument and to diagnose specific problems, in this case, with how raters use scales to make cross-linguistic comparisons of phonological and semantic similarity. This section is organized in terms of the two analyses followed by a general discussion.

Phonological similarity ratings

The results support the use of fewer rating categories for cross-linguistic phonological similarity. This finding is important for the general research community, where typically a 7-point scale has been used. In the present study, the lower rating categories were seldom used, with the very sparse data misfitting the Rasch model. The simulated 4-point scale, with the lowest categories collapsed into a single category, showed improved data-model fit. However, rater behavior may change

in unexpected ways if the number of categories is reduced, so empirical replication with a 4-point scale is required to confirm the finding of improved data-model fit.

Perhaps the primary concern of researchers in determining the appropriate length of the scale is to maximize data variation while not restricting it by including too few categories. Hence, a 7-point scale, even if part of it is not fully utilized, provides a wide range of responses while not overly restricting them. Moreover, while scales can be collapsed after data collection (e.g., in Miwa et al., 2014), they cannot be expanded. Hence, for practical purposes, adopting a 7-point scale initially would seem a prudent decision. The alternative argument, however, concerns the theoretical construct that is represented through the use of the scale. If raters cannot visualize certain parts of the scale well, it suggests that humans perhaps perceive less variation than that conceptualized in the scale. In this case, there becomes a theoretical basis for removing redundant categories so that they better match human perception, which should, as shown here, improve the objective measurement of the construct. Researchers must decide on the length of the scale based on these practical and theoretical concerns, yet our study shows that a compromise of the two may be needed, and that a 7-point scale is perhaps too much for cross-linguistic similarity ratings, at least for Japanese-English cognates and when noncognates are excluded. If studies include noncognates, then they will occupy the most dissimilar point on the scale, as shown in previous studies (i.e., Allen & Conklin, 2014; Tokowicz et al., 2002), and thus an additional point on the scale may be needed for these 'completely different' items (i.e., a 5-point scale).

Regarding variation across persons and items, one important finding was of substantive rater disagreement. The person reliability coefficients exceeded .98, with a large logit range of severity, indicating that raters cannot be considered interchangeable in terms of severity. In particular, three raters behaved idiosyncratically, evidenced by the fit analysis. This idiosyncratic behavior is relative to the average rater, meaning that it is sample dependent. Nevertheless, it raises a number of important questions that researchers must consider in the development of research instruments. Firstly, is the variation systematically related to rater characteristics, such as language proficiency? If so, these characteristics must be controlled when recruiting raters. In the present case, raters did not differ substantially in terms of lexical knowledge (as measured using the *Vocabulary Size Test*), which rules it out as a possible explanation. It is of course possible that another rater characteristic is systematically related to the variation in ratings, yet it is not clear what this could be.

Another question raised is whether to exclude misfitting raters from subsequent analyses. Here, we agree with Davidson (2000) that excluding raters on an ad hoc basis is a problematic response to issues of misfit. It is better to attempt to identify the reasons for rater misfit, which will allow for more principled rater selection in future studies. In addition, misfitting raters highlights the importance of sample size in data collection. Researchers have varied considerably in the number of raters that they have recruited for cross-linguistic similarity rating studies, yet the results presented here illustrate the possibility that idiosyncratic behavior will be observed, suggesting that researchers apply caution and collect data from multiple raters for each item. In the present study, item difficulty was extremely stable due to the relatively high number of raters, indicating that useful conclusions can made on the basis of the rating data.

Moreover, the item dependency analysis raised our awareness of an important issue connected to the purpose of the instrument. In Allen et al. (2021), the purpose of the rating study was to determine the perceived phonological similarity of specific word pairs that would be used in a subsequent task, that is, self-paced reading. However, if the purpose was to develop a rating instrument for understanding perceived phonological similarity more generally, that is, where ratings were generalizable to other word pairs not included in the instrument, then items must be selected according to their specific characteristics relative to other items. Our results suggested that eliminating dependency may require redefining 'item' to apply to phonological features rather than words themselves. For instance, words that are highly similar, such as *wind*-ウインド /ɯndo/ and *wing*-ウイング /ɯngu/, actually represent two instances of the item *winX*-ウインX. In other words, these two word-pairs share all but the word-final phoneme, which is converted relatively consistently into *katakana* (i.e., -*d* to ド /do/, and -*g* to グ /gu/, respectively). In terms of making practical research instruments, phonologically similar word pairs, such as *wind*-ウインド and *wing*-ウイング, could be administered in different test forms (i.e., to different subsets of students), which could then be linked using Rasch analysis. If the dependency was high, the two words could be combined into a single item based on the shared phonological feature, but they could be treated as two separate items if the dependency was low. In this way, Rasch analysis provides opportunities not only for insights into rating behavior but also for the creation of useful research instruments.

In response to the third research question, the logit measures represented a near-linear transformation from the raw ratings, for both the 7-point scale and the 4-point scale. Although this result cannot be assumed to generalize to all datasets, it provides evidence that the raw ratings from this dataset can be validly interpreted as measures of the phonological similarity of the loanwords.

Semantic similarity ratings

The results show that raters were unable to effectively distinguish seven levels of semantic difference, with improved psychometric results from rescoring the data as dichotomous responses. Although empirical confirmation is required, the answer to the first research question is that a dichotomous scale appears to be optimal. Importantly, this is for cognate word pairs; if noncognate word pairs are included (e.g., *wall*-テーブル /te:buru/ 'table'), they will occupy the most dissimilar point on the scale and therefore a 3-point scale would be the minimum size, increasing to a maximum 5-point scale. This also applies to false friends, which share form but differ in meaning. Overall, the finding that semantic similarity of cognates is perhaps best measured on a shorter scale is of importance for researchers who utilize measures of cross-linguistic similarity in their work. This also agrees with the decision made by Miwa et al. (2014), who collapsed semantic ratings to a dichotomous (i.e., identical, non-identical) scale. Future studies should demonstrate more conclusively whether this is recommended more generally for research in this area.

In response to the second research question, while there was some variation in responses, by-and-large the raters performed consistently, with a tendency to rate word pairs as being very similar across languages. Some misfitting items were observed, but it was not possible to isolate the effects of items from raters. Three raters showed concerning levels of misfit, two of whom were misfitting in the phonological similarity analysis, suggesting that these raters had some background characteristic that made them rate differently from the group. As discussed previously, this was not lexical proficiency. One possibility is that these raters were not performing the task correctly, perhaps due to difficulties in staying on task. However, this conjecture cannot be supported on the basis of the data. To investigate such issues, utilizing interviews and retrospective think-aloud methodology would undoubtedly shed some light on the actual reasons behind such idiosyncratic behaviors. We leave this suggestion open for future studies.

A more general issue connected to the variation in responses for semantic similarity is how best to measure it. Previous researchers in applied linguistics have tried to categorize loanwords according to their formal and semantic characteristics. For example, Uchida (2007) categorized Japanese loanwords as *true cognates, convergent cognates, divergent cognates, close false friends, distant false friends* and *Japanised English*. This method of categorizing loanwords is fraught with difficulties, however, even for the linguistic expert, never mind the typical language learner. Consequently, the validity of such an approach is compromised, necessitating a more valid and practical approach to the measurement of the cross-linguistic similarity of loanwords. Based on the research presented here, we advocate the use of formal and semantic scales, and that Rasch analysis can be used in the development of these scales.

In response to the third question, the relationship between raw scores and logit measures was not strongly linear due to many items being judged to be extremely similar. Scores at the extremes of the range are inevitably distorted, so logit measures are preferable in this instance. However, raw scores reflected the ordinal ranking of item difficulty (i.e., similarity), which may be sufficient for many research purposes.

Limitations and future directions

Although there is a wealth of robust evidence within psychology for the cognate effect, it must be noted that the effect is typically small. The cognate effect is typically revealed as an imperceptible average advantage in word reading of around 50 milliseconds (i.e., 50 thousandths of a second). Therefore, while language learners may perceive some words to be easier to recall, produce or learn, the extent of the cognate effect in everyday language use typically goes unnoticed. The implication of this for research in applied linguistics and classroom research is that subtle differences in phonological and/or semantic similarity may not appear to make much difference in terms of learners' language use. Rather, it may be that the benefits of cognates are in fact much less obvious relative to the often-disruptive effects of words that share form but not meaning (i.e., false friends, false cognates, homophones, and homographs). In psycholinguistic terms, the co-activation of similar formal features across languages leads to activation of competing semantic representations, which slows down processing. For instance, *consent* in English would activate コンセント /konsento/ in Japanese, which has both the same meaning as the English word, as in 'informed consent', but also a different meaning, as in 'electrical outlet'. This disruption in processing is likely to be observed in classroom research, especially for words that are maximally different in meaning across languages. Moreover, cognates that are very different phonologically across languages, such as *varnish*-ニス /nɪsu/ 'varnish', are most likely not to benefit from cross-linguistic similarity. However, for all of the thousands of words that do share considerable overlap in form and meaning, while there may be a benefit conveyed to learners, this benefit will often go unnoticed.

As described previously, future studies should seek to empirically validate shorter scales for cross-linguistic similarity. Such validation studies should utilize Rasch analyses, but also qualitative data of participants' explicit thought processes, as monitored through a think-aloud protocol, would usefully supplement the ratings data. Taken together, these different

sources of information can be used to make decisions about the optimum instrument for measuring cross-linguistic similarity.

Acknowledgements

We would like to thank the two reviewers for their helpful feedback.

Notes

[1] It should be noted that Miwa et al.'s (2014) scale was the reverse of that typically used; that is, a rating of '1' indicated 'identical', whereas in most other studies '1' indicated 'completely different'. However, it is very unlikely that this mirroring of the rating scale would seriously impact the measure.

References

Allen, D. (2019a). Cognate frequency and assessment of second language lexical knowledge. *International Journal of Bilingualism, 23*(5), 1121–1136. https://doi.org/10.1177/1367006918781063

Allen, D. (2019b). Cognate frequency predicts accuracy in tests of lexical knowledge. *Language Assessment Quarterly, 16*(3), 312–327. https://doi.org/10.1080/15434303.2019.1635134

Allen, D. (2019c). The prevalence and frequency of Japanese-English cognates: Recommendations for future research in applied linguistics. *International Review of Applied Linguistics in Language Teaching, 57*(3), 355–376. https://doi.org/10.1515/iral-2017-0028

Allen, D., & Conklin, K. (2013). Cross-linguistic similarity and task demands for Japanese–English bilingual processing. *PLoS One, 8*(8), e72631. https://doi.org/10.1371/journal.pone.0072631

Allen, D., & Conklin, K. (2014). Cross-linguistic similarity norms for Japanese-English translation equivalents. *Behavior Research methods, 46(2),* 540-563. https://doi.org/10.3758/s13428-013-0389-z

Allen, D., Conklin, K., & Miwa, K. (2021). Cross-linguistic lexical effects in different-script bilingual reading are modulated by task. *International Journal of Bilingualism, 25*(1), 168-188. https://doi.org/10.1177/1367006920943974

Andrich, D. (1978). A rating formulation for ordered response categories. *Psychometrika, 43*(4), 561-573. https://doi.org/10.1007/BF02293814

Aryadoust, V., Ng, L. Y., & Sayama, H. (2021). A comprehensive review of Rasch measurement in language assessment: Recommendations and guidelines for research. *Language Testing, 38*(1), 6-40. https://doi.org/10.1177/0265532220927487

Davidson, F. (2000). The language tester's statistical toolbox. *System, 28*, 605-617. https://doi.org/10.1016/S0346-251X(00)00041-5

DeMars, C. (2010). *Item response theory*. Oxford University Press.

Dijkstra, T., Grainger, J., & van Heuven, W. J. B. (1999). Recognition of cognates and interlingual homographs: The neglected role of phonology. *Journal of Memory and language, 41*, 496-518. https://doi.org/10.1006/jmla.1999.2654

Dijkstra, T., Miwa, K., Brummelhuis, B., Sappeli, M., & Baayen, R. H. (2010). How crosslinguistic similarity affects cognate recognition. *Journal of Memory and Language, 62*(3), 284–301. https://doi.org/10.1016/j.jml.2009.12.003

Dijkstra, T., & Van Heuven, W. J. B. (2002). The architecture of the bilingual word recognition system: From identification to decision. *Bilingualism: Language and Cognition,* 5(3), 175–197. https://doi.org/10.1017/S1366728902003012

Dijkstra, T., Wahl, A., Buytenhuis, F., Van Halem, N., Al-Jibouri, Z., De Korte, M., & Rekké, S. (2019). Multilink: a computational model for bilingual word recognition and word translation. *Bilingualism: Language and Cognition,* 22(4), 657-679. https://doi.org/10.1017/S1366728918000287

Engelhard, G. (2013). *Invariant measurement*. Routledge.

Hoshino, N., & Kroll, J. (2008). Cognate effects in picture naming: Does cross-linguistic activation survive a change of script? *Cognition, 106*(1), 501–511. https://doi.org/10.1016/j.cognition.2007.02.001

Linacre, J. M. (1994). Many-facet Rasch measurement. (2nd ed.). MESA Press.
Linacre, J. M. (2009). *Misfit diagnosis: infit outfit mean-square standardized*. http://www.winsteps.com/winman/index.htm?globalfitstatistics.htm
Linacre, J. M. (2016). *Dimensionality investigation - an example*. http://www.winsteps.com/winman/multidimensionality.htm
Linacre, J. M. (2020). *Table 23.99 Largest residual correlations for items*. https://www.winsteps.com/winman/table23_99.htm
Linacre, J. M. (2020). Winsteps (Version 4.6.2) [Computer Software]. Winsteps.com.
Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika, 47*(2), 149-174. https://doi.org/10.1007/BF02296272
McNamara, T. F. (1996). *Measuring second language performance*. Pearson Education.
Miwa, K., Dijkstra, T., Bolger, P., & Baayen, H. (2014). Reading English with Japanese in mind: Effects of frequency, phonology, and meaning in different-script bilinguals. *Bilingualism: Language and Cognition, 17*(3), 445–463. https://doi.org/10.1017/S1366728913000576
Nation, I. S. P., & Beglar, D. (2007). A vocabulary size test. *The Language Teacher, 31*(7), 9–13. https://jalt-publications.org/files/pdf/the_language_teacher/07_2007tlt.pdf
Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Denmark Paedogiske Institut.
Reckase, M. D. (1979). Unifactor latent trait models applied to multifactor tests: Results and implications. *Journal of Educational and Behavioral Statistics, 4,* 207-230. https://doi.org/10.3102/10769986004003207
Schepens, J., Dijkstra, A., & Grootjen, F. (2012). Distributions of cognates in Europe as based on Levenshtein distance. *Bilingualism: Language and Cognition, 15*, 157–166. https://doi.org/10.1017/S1366728910000623
Stevens, S. S. (1946). On the theory of scales of measurement. *Science, 103*(2684), 677-680.
Tokowicz, N., Kroll, J. F., De Groot, A. M. B., & Van Hell, J. G. (2002). Number-of- translation norms for Dutch–English translation pairs: A new tool for examining language production. *Behavior Research Methods, Instruments, & Computers, 34*(3), 435–451. https://doi.org/10.3758/BF03195472
Van Assche, E., Duyck, W., Hartsuiker, R. J., & Diependaele, K. (2009). Does bilingualism change native-language reading?: Cognate effects in a sentence context. *Psychological Science, 20*(8), 923–927. https://doi.org/10.1111/j.1467-9280.2009.02389.x
Van Orden, G. C. (1987). A rows is a rose: Spelling, sound, and reading. *Memory & Cognition*, *15*, 181–198. https://doi.org/10.3758/BF03197716
Wright, B. D., & Stone, M. H. (1979). *Best test design*. MESA Press.

The Construction and Validation of a New Listening Span Task

Bartolo Bazan
bazanlinkin2@gmail.com
Department of English, Ryukoku University Heian Junior & Senior High School
https://doi.org/10.37546/JALTSIG.TEVAL25.1-4

Abstract

The listening span task is a measure of working memory that requires participants to process sets of increasing numbers of utterances and store the last word of each utterance for recall at the end of each set. Measures to date have contained an exceedingly demanding processing component, possibly leading to insufficient resources to meet the word recall requirement, which may have affected the sensitivity of the measure to distinguish different levels of working memory. Further, tasks thus far have asked participants to verify the content utterances based on knowledge, which may have confounded the measurement of working memory capacity with world knowledge. An additional weakness is that they lack sound psychometric construct validity evidence, which clouds what these tools actually measure. This pilot study presents a listening span task that accounts for preceding methodological shortcomings, which was administered to 31 Japanese junior high school students. The participants listened to ten sets (two sets of equal length of two, three, four, five and six utterances) of short casual utterances, judged whether they made sense in Japanese, and recalled the last word of each utterance in the set. Performance was assessed through a scoring procedure new to listening span tasks in which credit is given for the words recalled in order of appearance until memory failure. The data was analyzed through the Rasch model, which produces evidence for different aspects of validity and indicates if the items in a test measure a unidimensional construct. The results provided validity evidence for the use of the new listening span task and revealed that the instrument measured a single unidimensional construct.

Keywords: working memory, listening span task, validation, Rasch model, Japanese

Working Memory (WM) refers to a mental workspace where information, retrieved from long-term memory, is held and simultaneously manipulated (Baddeley et al., 2002). WM has been closely associated with performance on a wide range of cognitive skills such as first and second language use (Gathercole & Baddeley, 1993; Linck et al., 2014).

A number of complex span tasks, which are tasks that tap both the processing and storage functions of WM, have traditionally been employed to measure WM capacity. An example of such tasks is the listening span task, which is the focus of this study, and in which an individual is asked to verify the grammaticality of each of a series of utterances at the same time as retaining the sequence of the final words of preceding utterances (Daneman & Carpenter, 1980). However, despite the general acceptance of these tasks as measures of WM capacity, few efforts have been made to revise the tasks and/or to collect validity evidence for their use. This is problematic for three reasons. First, the theoretical premise underlying complex WM tasks implies that individuals with efficient processing capacity will also have larger storage capacity (Daneman & Merickle, 1996). Thus, WM tasks, including the listening span task, have been designed to measure WM storage with the interference of perhaps excessively demanding processing components. If test-takers need to maximally employ their available resources to meet the task processing requirement, they may have insufficient capacity to temporarily store target items. This has resulted in a narrow spread of item-recall scores, which suggests that the tasks may not be sensitive enough to differentiate people with different WM levels. Second, the tasks have not been developed to account for confounding methodological factors such as potential knowledge biases in the utterance verification component of listening span tasks. That is, tasks that involve judging the plausibility of utterances based on knowledge may measure both WM span and general knowledge and may therefore provide inaccurate WM measurement. Third, the tasks' construct validities have not been well supported as the available validity evidence has been limited to people with frontal lobe lesions (Miyake et al., 2000) as well as the fact that WM instruments predict performance on a wide range of tasks, such as following directions, note-taking, reading, and writing (Conway et al., 2005). It is thus unclear what WM tasks actually assess.

Based upon these three issues, it is fair to state that there is a paucity of both rigorous methodological revisions of WM tasks as well as psychometrically sound validity evidence, such as that provided by the Rasch model (Rasch, 1960). The goal of this pilot study is to (a) develop a complex span task, namely a listening span task, which addresses the flaws of its predecessors, as well as to (b) collect validity evidence for its use through Rasch analyses. Rasch analysis provides detailed information about different aspects of validity and can reveal whether a set of items is functioning to measure a single underlying construct, such as WM capacity. In this paper, validity refers to the strength of evidence in support of inferences about a human trait that can be made from an observed performance on a task. Validity is assessed by analyzing the person and item fit and the person and item reliability and separation (Bond & Fox, 2015). The results of this pilot study will be

used as a baseline to develop a computerized listening span task that can be administered to multiple participants simultaneously rather than individually as in the current task procedure.

The Listening Span Task

The listening span task is a complex span task that was originally developed by Daneman and Carpenter (1980) to gauge WM capacity, which is the capacity to store information with the interference of processing demands. The task was constructed with 60 utterances of between nine and 16 words in length which had been taken from quiz books and whose content covered a variety of knowledge domains such as literature, biological sciences, and geography. To illustrate, one of the utterances participants heard was *You can trace the languages English and German back to the same roots* (p. 458). The participants were required to listen to 15 sets of utterances (three each composed of two, three, four, five, and six utterances, respectively) and hold in memory the final word of each utterance after having judged the truthfulness of the statement. Half of the utterances were true, while the other half were false. Participants had to decide whether the presented utterance was true or false and were given a second and a half to attempt to store the last word so that they did not have time to rehearse the words in their minds. If participants did not verify the utterance within the time given, they were pushed to answer quickly or, if they did not know the answer, they were presented with the next utterance. The true-or-false component was added as a distractor and was not included in the score. Upon completion of the set, participants heard a beep signaling that they could start to recall the final words in the sets in order of appearance. The participants' WM span was defined as the utterance-level at which they were correct on two of the three sets. If the participants recalled all the words of only one set of the three, they were awarded half a credit. The test was finished when the participants could not recall any of the words in all three sets at a particular level. For example, if a participant was correct on two sets at the two-, three-, and four-utterance-levels but was incorrect on all three sets at the fifth level, they were given a span of 4.00. If the participants performed correctly on one of the three five-level utterance sets, they were given a score of 4.50. However, due to the fact that the task was excessively demanding on the participants, as Daneman and Carpenter themselves acknowledged, a credit was given for any set at which all of the words were recalled, regardless of their order of appearance.

Drawing on Daneman and Carpenter's (1980) test specifications, Osaka et al. (2003) designed a Japanese version that required participants to hear three sets of four utterances, judge their semantic plausibility, and store in memory the first word of each utterance in the set for oral recall at the end of the set. The utterances were six seconds long and were presented at one second intervals. The rationale for using the initial word in the utterances was that utterances tended to finish with a verb in Japanese. The high performing participants obtained a mean word recall score of 96.90 whereas the low-performing participants' mean score was 89.10. However, because an explanation of the scoring procedures was not reported, it is unclear how these average scores were computed. In addition, example utterances were lacking.

Another Japanese version of the task constructed by Komori (2016) contained 70 utterances between 35 and 46 mora[1] long (M = 41.77) that were divided into five sets of two, three, four, and five utterances. Similar to Daneman and Carpenter's (1980) listening span task, Komori's task required participants to listen to the increasingly longer sets and judge whether the utterances were true or false based on general knowledge. Half of the utterances were true and the other half were false. Simultaneously, as in Osaka et al.'s (2003) task, the participants had to remember the initial word (a noun) of each utterance and recall them at the end of each set. Although participants were allowed to recall the target words in any order, they were not allowed to start the recall with the last target word in the set. An example set of two utterances is as follows: (1) *migi tede chokiwo tsukuri hidari tede paawo tsukuruto orarete iru teno yubiwa nihonto naru* [When you make scissors with the right hand and paper with the left hand while playing rock-paper-scissors, the number of folded fingers is two] and (2) *denwawa onseiwo shingou henkashite hanareta aiteni tsutaeru monode, keitai gatamo aru* [Telephones are devices that encode voices as signals to communicate with a distant person, and include mobile phones] (Komori, 2016, p. 4).

The scoring system utilized by Komori (2016) was similar to that used by Osaka et al. (2003), and WM capacity was calculated as the maximum set size at which the participants could recall all the words in three of the five sets. An additional half credit was given if the participants were successful on two of the following difficulty sets. For example, if participants recalled all of the words in three sets of two utterances and two sets of three utterances (the following difficulty level), they were awarded 2.50 points. An inspection of the descriptive statistics table, however, revealed that the task was difficult for the sample. A subgroup of participants, classified as high spans, obtained a mean word recall score of 0.96 (SD = 0.08), 0.96 (SD = 0.06), 0.91 (SD = 0.07), 0.84 (SD = 0.07) for the sets of two, three, four, and five utterances, respectively. In other words, they recalled on average less than one word per difficulty level.

In fact, the three listening span tasks reviewed above all seemed to be highly demanding due to the length of the utterances found therein, which not only affected the amount of information that needed to be held in memory for processing, but also increased the duration of the retention interval, possibly resulting in the decay of the words temporarily stored in memory (Towse, et al., 2000). Thus, if the listening span task requires the participants to allocate an excessively large amount of resources to the processing component, it is likely that they will be left with insufficient storage capacity to meet the item-recall component requirement effectively. There is some evidence to support this hypothesis. In a study that compared individuals with and without aphasia under different WM conditions, Ivanova and Hallowell (2014) found that long utterances negatively impacted word-recall by the non-aphasic participants as opposed to the aphasic participants.

Furthermore, a highly demanding test would yield a narrow spread of scores, as occurred in Komori's (2016) study, which in Rasch measurement would translate into low item and person reliability. According to Bond and Fox (2016), this is because the test-takers with lower WM capacity would not have items that targeted their WM level, whereas the hardest sets of items would not allow test-takers with sufficient WM capacity to provide information about their functioning. Consequently, the person ability separation index would be low, which is an indication that the task may be insufficiently sensitive to distinguish people with different WM levels.

An additional methodological limitation is that the true/false verification component of previous listening span tasks was based on general knowledge, thus confounding WM performance with world knowledge. That is, knowledgeable test-takers may have scored higher and less knowledgeable test-takers lower than would be expected in a WM measure without this knowledge bias, which suggests that previous listening span tasks have provided an imprecise measurement of WM capacity. Insofar as the true/false component entails a judgement based on knowledge, the listening span task would measure both WM capacity and knowledge rather than true WM capacity. Thus, while the original authors did attempt to control for knowledge by selecting content that would be likely known to all potential test-takers, it is still possible that it could impact performance.

Lastly, along with these task-requirement limitations, the scoring methods utilized by the authors of the preceding studies allowed the participants to recall the target words in any order. This may have impacted the hypothesized hierarchy of difficulty of the items (because the further in the set that the words appeared, the more difficult the item should be to recall). For example, as participants could begin by recalling the hypothesized most difficult items (the last items in the sets), the difficulty level of those final items would fall below the difficulty level of the preceding items in the set, which are theorized to be easier. The present study has been designed to address these weaknesses.

Research Questions

The goal of the current study was to address the weaknesses of previously published listening span tasks by developing a new task and collecting validity evidence for its use through Rasch analysis. With this in mind, the present study was guided by the following research questions:

1. Do the items within the new listening span task (NLST) sets gradually increase in difficulty as hypothesized (i.e., the further their position within the set, the more difficult they should be)?
2. Does the NLST data fit the Rasch model?
3. Is the NLST item reliability sufficient to suggest replicability of the item difficulty hierarchy if the listening span task is administered to a similar sample?
4. Is the NLST person reliability sufficient to suggest a similar spread of participants with higher and lower levels of the construct (WM capacity) if the same participants were administered a similar item sample?
5. Does the NLST separate the assessed participants into different levels of the construct (higher and lower WM capacity)?
6. Is the NLST unidimensional?

Each Research Question addresses different aspects of construct validity (Bond & Fox, 2015). Together, the findings for each of these questions provide evidence towards validity claims for the listening span task.

Methodology

Participants

This study took place at a private junior and senior high school in Western Japan. At this institution, students were streamed into classes by academic level, comprising high-, intermediate-, and low-level classes. The 31 participants who performed the listening span task came from the second grade of junior high school and were aged 13 and 14 years old. The sample was composed of nine volunteer students from the low-level class, 10 from the intermediate class, and 12 from the high-level class. There were 20 female students and 11 male students, who were more or less equally distributed among the three levels. All participants were native speakers of Japanese.

Instruments and Administration

A shortened listening span task, which consisted of 40 unrelated casual utterances about daily-life situations (see Appendix A), was developed for the purposes of this study. The test differed from previous versions in several ways. First, utterance length was shorter with a range of between three and five words. This modification was made to keep the processing component of the task from exceeding WM capacities. Second, the task contained fewer items (20 less than the original version and 30 less than Komori's [2016] test) and there was no practice session prior to the test. The rationale for these changes was that the longer the task, the more likely it would be that participants would engage in idiosyncratic strategies to complete it (Miyake et al., 2000), which may confound measurement. Further, longer tasks have the disadvantage that participants may become tired, which would negatively impact their performance. Additionally, practice trials were not included because, unlike other span tasks such as the Tower of Hanoi or the Wisconsin sorting test (Miyake et al., 2000), listening span measures are relatively simple. Also, the task was not computerized because this step may have required additional trials for the participants to become familiar with the functions of the keys.

Third, not only did the NLST test items differ from those in previous WM tests in terms of length and number of items. They also differed in terms of the content verification component and target words for recall. Instead of a content verification component based on knowledge, this study used a grammaticality judgement test that required participants to verify if the utterances made sense in Japanese, which may have accounted for the knowledge bias of previous versions. Half of the utterances were grammatical and the other half ungrammatical (incorrect word order) and they were randomly arranged into two sets of two, three, four, five, and six utterances. Fourth, in contrast to its Japanese predecessors, the utterance-final words served as the to-be remembered items. This change had the benefit of reducing the level of memory decay as the duration of target-word retention was minimized by its final position in the utterances, thus reducing the likelihood of reliance on verbal rehearsal strategies to recall the words. Although the reason for previous listening span measures to use the initial words as target words was that Japanese utterances tend to end in verbs, which may make recall easier, casual Japanese utterances can also end in adjectives. Furthermore, ungrammatical utterances do not need to end in verbs. In this task, the target words were 12 adjectives, 11 nouns, 11 verbs, three quantifiers, and three adverbs, most of which were two or three mora long (see Appendix A). Further, an attempt was made to control for the complexity and frequency of the words, by including only words that the researcher deemed easy and highly frequent. Two students at the same institution who were unrelated to the study confirmed that all words were morphologically simple and likely to be known to participants. The utterances were audio-recorded by a female Japanese native speaker. The task was preceded by a demonstration of how to perform it.

The shortened listening span task was administered by the author of the present study one-on-one in a quiet room and was conducted entirely in Japanese. Before the test began, the participants received written and oral instructions that asked them to judge the grammaticality of each utterance and recall the final word in each utterance in the set in the correct order at the end of the set. It was explained that if the utterance's final word had a particle attached to its end (i.e., *kireida*), participants did not have to recall the particle. Similarly, if the utterance ended in a verb inflected in the past tense, participants could recall it in its plain form. After the instructions, the participants had the opportunity to ask clarification questions.

The audio stimuli were presented by the author using Windows Media Player on a laptop computer and the test began with the two sets of two utterances, gradually moving up to the two sets of six utterances. Immediately after each utterance, the audio was paused and the participants judged if the utterance made sense in Japanese. At the end of each set, participants recalled the to-be-remembered words in their order of presentation. Each participant's performance was audio-recorded for scoring. The scoring procedure was adopted from Bazan (2020) and consisted of giving a credit for each word recalled in

a string in the correct order of appearance until memory failure to recall in order. For example, if on a set of five utterances, participants correctly recalled the last word of the first and second utterances, failed to recall the last word of the third utterance (i.e., *sanbanmewo oboete nai* [I don't remember the third word]), but succeeded in recalling the last word of the fourth and fifth utterances, they were given two points (one for utterance 1 and another for utterance 2) and the rest of their responses did not count.

The utterance verification component was used as a distractor to make sure participants processed the utterances, and thus was not scored (participants were not made aware of this information). This scoring system had the added advantage of preventing participants from benefiting from recency effects as participants had to recall the words in the exact order of appearance rather than in free recall. This design maintained the hypothesized order of difficulty of the words (i.e., the further in the set, the more difficult they should be to recall). Contrary to previous scoring procedures (Friedman & Miyake, 2005), where the test was terminated when participants failed to perform perfectly on a particular set, participants in this study were administered all of the sets, regardless of how many words they could recall, which gave participants equal opportunities.

Rasch Analysis

The data were entered into a spreadsheet, which was imported to Winsteps 4.3.1 (Linacre, 2018) for analysis using the Rasch dichotomous model (i.e., items scored as right or wrong). Research Question 1 was addressed by an examination of the Wright map. Research Question 2 was answered by looking at the item and person fit indices. Research Questions 3, 4, and 5 were explored using the person and item reliability and the separation indices, respectively. Research Question 6 involved a principal components analysis (PCA) of item residuals and an inspection of the item fit graph. These are dimensionality indicators about whether the test assesses a single construct.

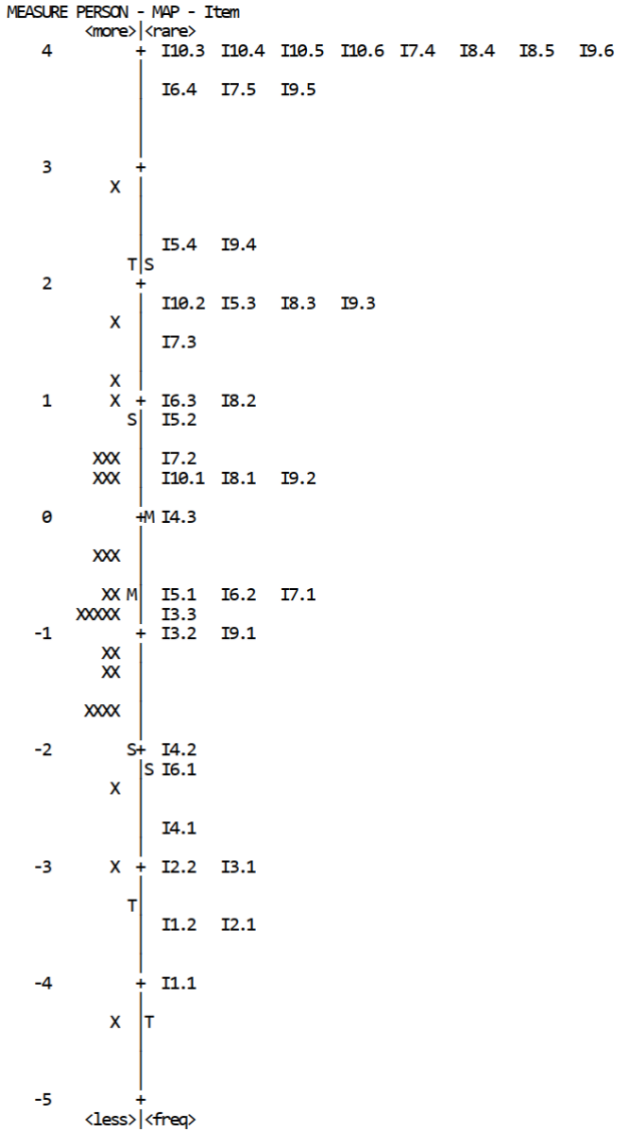
Results

Item-Person Map or Wright Map

Figure 1 shows the item-person relationships plotted on the map of the listening span task. The line in the center of the map represents the distances between points on the logit scale, which is an interval scale measurement, and which is shown numerically on the left. In other words, the distances between the data points along the line are thought to represent equal amounts of WM capacity. The participants, who are each represented by an 'X', are located on the left side in ascending order of hypothesized WM capacity. That is, the higher up the map, the higher the participants' scores on the listening span task. Similarly, the items are spread on the right side in ascending order of difficulty. In other words, the higher up the map, the more difficult the item. The plot shows that the individual items within each set are ordered in accordance with the theoretical expectation that the later the item appears in the set, the more difficult it should be. For example, the last item of set 3 (item 3.3) is higher than the second (item 3.2), which is higher than the first (item 3.1). As shown, a number of items in sets 7, 8, and 10 (items 7.4, 7.5, 8.4, and 8.5 and items 10.3, 10.4, 10.5, 10.6, respectively) do not align with the theorized difficulty hierarchy. For example, items 10.3, 10.4, 10.5, and 10.6 are shown to be equally difficult. This is explained by the fact that the fit statistics of those items were not estimated by Winsteps as no participant was successful on them. An alternative explanation is that there were not enough participants at this level to discriminate the difficulty hierarchy of the more difficult items. The distribution of participants is heavier in the lower half of the figure (below the 0.00 logit measure to the left of the persons, representing the mean item difficulty), which suggests that a greater number of higher WM-span participants were needed. The participants were, however, well spread out over approximately seven logits along the WM logit scale.

Figure 1

Wright map for the listening span task individual items analysis



Note. "X" represents each individual participant's performance, "I" are the items, which are followed by the set number and the item number, the logit scale is all the way on the left, under Measure. The line down the middle separates items and persons, locating these facets on a common frame of reference in keeping with the Rasch model.

Person and Item Fit Statistics

The Rasch fit statistics are quality-control indicators that are useful to evaluate the degree to which the data meet the model's expectations. Rasch provides two aspects of fit, namely infit mean-square (MNSQ), which is a weighted unstandardized form of fit, and outfit MNSQ, which is a non-weighted standardized fit statistic that is sensitive to outliers (Linacre, 2002; Bond & Fox, 2015). As the outfit statistic is affected by outliers (unexpected performances of participants who manage to succeed on items above their abilities), infit MNSQ tends to be the statistic that guides the assessment of fit (Bond & Fox, 2015). In this study too, decisions about fit will be made based on infit MNSQ, but outfit values will also be examined to investigate unexpected performances of persons and items. Based on Linacre's (2002; 2007) guidelines, fitting persons and items were defined as those with infit MNSQ values of between 0.50 and 1.50 with perfect fit being indicated by a value of 1.00.

The person infit MNSQ indices for the participants in the listening span (see Table 1), revealed that all participants but one (participant c301, infit MNSQ = 1.81) had infit MNSQ values within the acceptable parameters of 0.50 and 1.50, indicating

that the sample behaved as expected by the model. The person infit MNSQ values, excluding person c301, ranged from 0.52 (person c102) to 1.44 (person c112), which shows that the participants' performance had acceptable fit.

The high infit MNSQ statistic for participant c301 (infit MNSQ = 1.81) is accompanied by a large outfit MNSQ value of 2.65, which suggests that this participant's performance was unexpected by the model. This was also true for participants c112, c136, and c229 who had large outfit MNSQ indices of 4.03, 3.83, and 2.89, respectively. An examination of their individual data revealed that participants c112, c136, and c301 were low performers (logit WM measures of -0.86, 0.53, and 0.66, respectively) who, perhaps through the use of an idiosyncratic strategy such as initial word mora recall or word chaining, managed to succeed on items above their WM level such as items 6.3 or 10.2 (difficulty measures of 1.00 and 1.86, respectively). In contrast, participant c229 was a capable participant (logit WM measure of 1.21) who unexpectedly failed on some easy items such as items 4.1, 4.2, and 4.3 (difficulty measures of -2.71, -1.96, and 0.00, respectively). The source, however, of these participants' poor outfit values seems to be the small size of the sample ($N = 31$) because a few unexpected responses can make the participants misfit in small samples (Boone & Noltemeyer, 2017).

Table 1

Person statistics for the listening span task

Person	Measure	*SE*	Infit MNSQ	Outfit MNSQ
c112	-0.86	0.6	1.44	4.03
c136	0.53	0.57	1.44	3.83
c229	1.21	0.55	1.29	2.89
c301	-0.62	0.66	1.81	2.65
c303	0.31	0.51	1.13	1.43
c125	0.98	0.57	1.42	1.13
c201	-1.12	0.58	1.32	1.06
c306	-1.66	0.61	1.31	0.93
c134	-0.86	0.57	1.29	1.06
c313	0.31	0.48	0.83	1.28
c225	-1.66	0.6	1.26	1.26
c307	-0.86	0.53	1.12	1.18
c114	0.53	0.5	1.12	0.85
c113	1.69	0.52	1.07	0.75
c321	-0.86	0.5	0.97	0.9
c102	0.31	0.48	0.93	0.69
c120	-0.86	0.5	0.89	0.67
c212	-2.25	0.56	0.85	0.53
c124	-1.38	0.52	0.85	0.49
c324	-4.34	0.8	0.79	0.21
c224	-1.38	0.52	0.76	0.5
c222	-0.62	0.49	0.75	0.51
c234	0.53	0.48	0.69	0.47
c103	-0.38	0.49	0.67	0.43
c108	-1.66	0.53	0.62	0.35
c314	-1.66	0.53	0.6	0.33
c210	-0.38	0.49	0.59	0.41
c203	-1.12	0.51	0.59	0.36
c209	2.84	0.58	0.58	0.28
c302	-0.38	0.49	0.53	0.34
c102	-2.94	0.61	0.52	0.21

Note. SE = standard error; MNSQ = mean-squared.

The item infit MNSQ values (see Table 2) showed a similar pattern of relatively well-behaved data. All values were within the cut-off parameters of 0.50 and 1.50 (i1.1 had the highest value, infit MNSQ = 1.42, and i6.1 had the lowest, infit MNSQ = 0.69), meaning that the items in the listening span task were successful in measuring the intended construct, thus providing evidence for construct validity (Bond & Fox, 2015).

The high outfit of items i1.1 (outfit MNSQ = 3.27), i4.1 (outfit MNSQ = 2.76), and i.6.4 (outfit MNSQ = 2.91) is explained by the fact that these items elicited unexpected performance by a few participants as an inspection of the Winsteps tables of item and persons responses revealed. For example, i1.1 had the lowest difficulty measure (-4.00 logits) and was supposed to be within all participants' WM capacities, yet two participants (c212 and c136) were unexpectedly unsuccessful on the item, which caused the high outfit (outfit MNSQ = 3.27).

Table 2

Item statistics for the listening span task

Item	Measure	*SE*	Infit MNSQ	Outfit MNSQ
i1.1	-4	0.96	1.42	3.27
i1.2	-3.45	0.68	1.01	1.08
i2.1	-3.45	0.68	0.93	0.72
i2.2	-3.04	0.6	0.97	0.67
i3.1	-3.04	0.69	1.34	1.12
i3.2	-1.03	0.41	0.85	0.75
i3.3	-0.86	0.41	0.97	0.94
i4.1	-2.71	0.61	1.25	2.76
i4.2	-1.96	0.48	1.09	1.59
i4.3	0	0.42	0.88	0.77
i5.1	-0.69	0.41	0.87	0.81
i5.2	0.77	0.47	0.86	0.75
i5.3	1.86	0.6	0.81	0.42
i5.4	2.26	0.67	0.76	0.37
i6.1	-2.19	0.48	0.69	0.49
i6.2	-0.69	0.47	1.31	1.25
i6.3	1	0.54	1.21	1.06
i6.4	3.63	1.2	1.27	2.91
i7.1	-0.69	0.41	0.99	0.91
i7.2	0.56	0.45	0.98	0.86
i7.3	1.53	0.55	0.82	0.68
i7.4	4.94	1.85	***	***
i7.5	3.63	1.18	1.22	0.93
i8.1	0.37	0.44	0.76	0.74
i8.2	1	0.49	0.96	0.94
i8.3	1.86	0.66	1.24	0.66
i8.4	4.94	1.85	***	***
i8.5	4.94	1.85	***	***
i9.1	-1.03	0.41	0.93	0.79
i9.2	0.37	0.44	0.96	0.95
i9.3	1.86	0.6	0.76	0.61
i9.4	2.26	0.71	1.14	0.87
i9.5	3.63	1.07	1.01	0.26
i9.6	4.94	1.85	***	***
i10.1	0.37	0.46	1.09	1.02
i10.2	1.86	0.6	1.01	1.07
i10.3	4.94	1.85	***	***
i10.4	4.94	1.85	***	***
i10.5	4.94	1.85	***	***
i10.6	4.94	1.85	***	***

Note. SE = standard error; MNSQ = mean-squared; *** = maximum measure.

Person and Item Reliability and Separation

The person and item reliability indicate the degree to which replicability of the person and item hierarchy is possible if the listening span test is given to a sample with similar characteristics. The higher the reliability value, the more confidence that can be placed in obtaining a similar ordering of persons and items across samples. The data revealed a person reliability estimate of .84 (the maximum possible value is 1.00), which is above the cut-off value of .80 (Linacre, 2007) and suggests that the probability of obtaining a similar spread of participants' WM capacities in similar samples is high. In addition, the person separation was estimated at 2.28 (see Table 3), which indicates that the sample was separable into three different levels of WM capacity (high, average, and low) as separation indices above 2.00 distinguish three distinct levels of the variable investigated (Duncan et al., 2003).

Table 3

Summary of the listening span task analysis (persons)

	Total Score	Count	Measure	Real *SE*	Infit MNSQ	Infit ZSTD	Outfit MNSQ	Outfit ZSTD
M	13.4	40	-0.6	0.55	0.97	-0.1	1.03	0.1
P. SD	5.2	0	1.37	0.07	0.33	1.2	0.98	1.1
S. SD	5.3	0	1.39	0.07	0.34	1.2	0.99	1.1
Max	27	40	2.84	0.8	1.81	2.6	4.03	2.9
Min	2	40	-4.34	0.48	0.52	-2.1	0.21	-1.4
REAL *RSME*	0.55	TRUE *SD*	1.25	SEPARATION	2.28	PERSON RELI.	0.84	
MODEL *RSME*	0.52	TRUE *SD*	1.27	SEPARATION	2.43	PERSON RELI.	0.86	
SE OF PERSON MEAN = 0.25								
PERSON RAW SCORE-TO-MEASURE CORRELATION = .99								
CRONBACH ALPHA (KR-20) PERSON RAW SCORE "TEST" RELIABILITY = .84								
SEM = 2.09								

Likewise, item reliability was .91 and item separation was calculated at 3.21 (see Table 4), meaning that the item difficulty hierarchy and spacing of items is highly replicable and that the listening span task separates items into four difficulty groups (Duncan et al., 2003). It is worth noting here that the total number of items provided in Table 4 (i.e., $N = 31$) is lower than the total number of items of the complete test (i.e., $N = 40$) because extreme scores are excluded. In any case, these results provide supporting evidence for construct validity (Bond & Fox, 2015).

Table 4

Summary of the listening span task analysis (items)

	Total Score	Count	Measure	Real *SE*	Infit MNSQ	Infit ZSTD	Outfit MNSQ	Outfit ZSTD
M	13	31	0	0.6	1.01	0	1.03	0.1
P. SD	9.3	0	2.15	0.21	0.19	0.7	0.68	0.6
S. SD	9.4	0	2.18	0.22	0.19	0.7	0.69	0.6
Max	29	31	3.63	1.2	1.42	1.9	3.27	1.8
Min	1	31	-4	0.41	0.69	-1.2	0.26	-0.8
REAL *RSME*	0.64	TRUE *SD*	2.05	SEPARATION	3.21	ITEM RELI.	0.91	
MODEL *RSME*	0.61	TRUE *SD*	2.06	SEPARATION	3.4	ITEM RELI.	0.92	
SE OF ITEM MEAN = .39								

PCA of Item Residuals and Item Fit Graph

A PCA analysis of the item residuals was conducted in order to examine the unidimensonality of the construct. There are two PCA requirements for unidimensionality. First, a unidimensional construct should account for 20.00% of the variance

or more (Reckase, 1979). Second, the principal contrast should produce an eigenvalue below 2.00 (Linacre, 2018) and explain less than 10.00% of the variance (Linacre, 2007). As shown in Table 5, the WM construct accounted for 51.30% (eigenvalue = 33.67) of the total variance, indicating that the instrument measured a single construct (the criterion value was at least 20.00%). Additionally, the principal contrast accounted for 6.50% of the unexplained variance, satisfying the unidimensionality criterion (< 10.00%). However, despite the fact that the first contrast explained less than 10.00% of the variance, its high eigenvalue (4.30) suggested the possibility of a second dimension.

Table 5

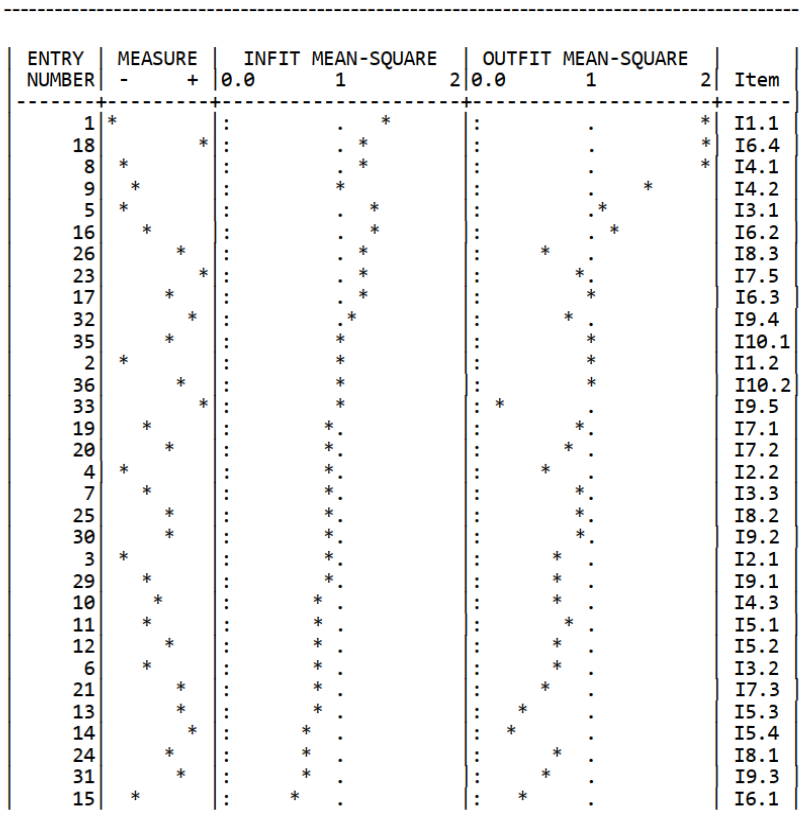
Listening span task standard residuals in eigen values

	Eigenvalue	Observed	Expected
Total Raw variance in observations	65.67	100.00%	100.00%
Raw variance explained by measures	33.67	51.30%	55.50%
Raw variance explained by persons	10.64	16.20%	16.00%
Raw variance explained by items	23.02	35.10%	34.60%
Raw unexplained variance (total)	32	48.70%	49.50%
Unexplained variance in 1st contrast	4.3	6.50%	
Unexplained variance in 2nd contrast	3.77	5.70%	

In any case, the item fit graph (see Figure 2) ruled out the possibility of an underlying second dimension. The infit MNSQ section of the figure depicts a unidimensional path. The straight vertical dotted line in the middle of the path represents the hypothesized unidimensional construct of WM capacity. As seen, the items of the listening span task, represented by asterisks, appear to be aligned along the ideal straight line. In addition, no item is outside the delimiting lines of the path, which would be of concern for the unidimensionality of the measure. Therefore, it seems reasonable to conclude that the listening span task is potentially unidimensional. However, a larger sample is required to make a clear conclusion.

Figure 2

Item fit graph for the listening span task



Discussion

The purpose of this study was to address the methodological issues of previous instruments by designing a listening span task and collecting sound psychometric validity evidence for its use through the Rasch model. It was argued that previous tests contained an exceedingly demanding processing component. The utterances in Osaka et al.'s (2003) instrument were six seconds long and those of Komori's (2016) ranged between 35 and 46 mora. In the latter study, the average recall rate was similar across sets (less than one word), which suggests that a set of two utterances was as difficult as a set of five. Additionally, the utterance verification of previous tasks such as found in Daneman and Carpenter's (1980) pioneering task, may have confounded WM measurement with world knowledge. The present listening span task addressed these shortcomings by controlling for utterance length and by having a grammaticality judgement test as the utterance verification component. Two additional new features were that the task lacked practice sets because the more practice, the more likely participants are to engage in strategies to complete the task (Miyake et al., 2000), and that it contained fewer items than its predecessors, which helps increase the practicality of its administration. Performance was scored adopting a scoring procedure that accounted for order of appearance to prevent participants from free recall, which is likely to involve strategic behavior.

Research Question 1 examined whether the items within the sets gradually increased in difficulty as expected based on theory (Daneman & Carpenter, 1980). An examination of the Wright map revealed that, overall, the difficulty of the items matched the theoretical expectations as the items were ordered along the map in ascending order of difficulty from initial to final set items. This means that, for example, recalling the target word of the third utterance in a set of three was more difficult than recalling the target word in the first utterance. This hierarchy of item difficulty is in contrast to that of Osaka et al.'s (2003) and Komori (2016). In these studies, the average recall rate was less than one item per set, which suggests that most items had a similar level of difficulty. The items in those studies may have been overly difficult, which may have impacted the precision of the measurement. These contrasting item difficulties can be explained by the impact of utterance length of the processing component of the task on the word storage component. The longer utterances used by Osaka et al. (2003) and Komori (2016) are likely to have produced greater interference and longer retention duration, potentially causing the to-be-remembered words to fade more easily. This explanation is in line with Cowan's (1999) embedded processes model of WM, which posits that WM is limited not only by the capacity to hold information but also, by the time the information can be held.

Research Question 2 asked whether the data fit the Rasch model. The majority of the items displayed good fit to the Rasch model, as none of the items had infit figures outside the established range (0.50 and 1.50). Three items (i1.1, i4.1, and i6.4), showed poor outfit (3.27, 2.76, and 2.91), but this was probably due to the unexpected performance of some participants, perhaps caused by a lack of concentration or nervousness at the beginning of the test. A most likely explanation for these high outfit values is, however, that no practice items were given to acclimate the participants to the task prior to its performance. This could have induced the participants to fail those items due to a lack of familiarity with the task procedures rather than due to a lack of ability.

Likewise, the participants performed close to the expectations of the model. One participant (c301) was identified as having large infit (1.81) and outfit (2.65) indices and three others (c112, c136, and c229) had large outfit indices (4.03, 3.83, and 2.39), which was explained by their off-target performance on several items probably due to the lack of practice trials. An alternative explanation is that the participants may have used idiosyncratic strategies such as initial mora recall to succeed on items that were above their level of ability.

All in all, the fit of the data to the Rasch model suggests that the construction of the task was, in general terms, effective. First, the grammaticality judgement task is likely to have served as a tool to make sure that the participants fully processed each utterance and that they did not simply focused on retaining the target words while ignoring the utterances (Turner & Engle, 1989). Second, although there are advantages to selecting the target words based on a corpus, such as a stricter control for word frequency and the elimination of a possible confound (i.e., the recall errors may be due to difficult word recognition rather than WM capacity), the intuitive approach used in the current study produced data that largely conform to the predictions of the Rasch model. Third, the results of the current pilot study lend support to the use of the scoring system corroborating the findings obtained by Bazan (2020). It is important to note, however, that this scoring system has the disadvantage that it requires participants to attempt all the trials (i.e., from Set 1 to Set 6), which may cause frustration once the task advances beyond the ability of the participants. In addition, this scoring system ignores one of the sources of the data, that of the processing component, as the grammatical verification of the utterances is not scored.

Research Questions 3 and 4 regarded the item and person reliability indices, respectively. These data revealed an item reliability coefficient of .91, which suggests that the spread of items along the WM continuum would likely be replicated if the NLST were given to another sample of similar characteristics. Similarly, the person reliability coefficient (.84) suggested high likelihood of reproducibility of the person hierarchy (Linacre, 2007). In other words, the participants would likely be placed at a similar level of ability if they were given a similar listening span task. These findings are consistent with the high reliability that complex span tasks have been shown to have based on split-half correlations or test-retest methods (Conway et al., 2005; Waters & Kaplan, 2004).

Research Question 5 investigated whether the listening span task separated the participants into different levels of ability. The participant performance was shown to be separable (separation = 2.28) into three levels of WM capacity (high, average, and low). This separation suggests that the top and bottom 23% of the sample had high and low ability, respectively, whereas the remaining 54% had average ability (Linacre, 2013).

According to Conway et al. (2005), the most common way of separating participants in the complex span task paradigm is quartile splits, in which the top and bottom quartiles of a distribution of WM scores are categorized as high and low span, respectively. This is the process of separation that both Komori (2016) and Osaka et al. (2003) used to split their respective samples for follow-up analyses. However, this process is problematic because it forces the separation groups to be equal in size, thereby treating participants, who may have different ability levels, as if they had the same ability level (Conway et al., 2005). For example, a group categorization based on quartile splits may give two groups of 30 participants each, but there may be different spreads of abilities within each group. The Rasch separation index is an alternative that is likely to yield a more precise separation and consequently more precise follow-up analyses. This is because the Rasch separation shows how many statistically differentiable ability levels exist in the population (Linacre, 2013), whereas quartiles might overestimate or underestimate the number of levels.

Research Question 6 explored the dimensionality of the measure. This dichotomous model explained 51.30% of the total variance (above 20.00%), which is one of the criteria of unidimensionality (Reckase, 1979). However, the results of the Rasch PCA contrast showed a concerning eigenvalue of 4.30, indicating the possible existence of a secondary dimension. Therefore, I examined the content of the items with the standardized residual loadings for the items in the Winsteps output to see if they hinted at a pattern (see Table 6). The items that were indicative of a possible subdimension seemed to share a common theme which could be called *familiarity* or *relevance to the participants' life*, as they were clearly broken into a cluster of utterances that had to do with the participants' everyday lives and a cluster of utterances that did not. For example, i3.2 *kouende tomodachito asobu* [I play with my friends in the park] vs. i9.4 *otousanwa inuga sukida* [My father likes dogs] or i6.3 *ekiga chikakattara benrida* [It is convenient to have the station close-by] vs i2.*1 kodomotachiga okashiwo kau* [children buy snacks]. This meant that the task could be confounded by the degree to which participants found the utterances related or not to their lives. There were, however, items that did not support this interpretation such as i6.1 *nihonwa supeinyori semai* [Japan is smaller than Spain] vs. i7.3 *utaimasu tomodachiga jouzuni* [sings my friend well], which is an ungrammatical utterance.

The possibility of this second dimension, however, was dismissed by the linear alignment of the items in the item fit graph. In general terms, these dimensionality results seem to provide support for unitary models of WM as opposed to models that consist of multiple separable subsystems (Miyake & Shah, 1999). Importantly, however, the sample size of this study is not large enough as to make a robust claim about the unitary versus the non-unitary nature of WM. Nevertheless, these results provide validity evidence for the measure and suggest new ways of developing and scoring listening span tasks.

Limitations

This study presents a number of limitations that should be addressed in future investigations. First, the sample size was too small to give sufficient statistical power to the results and therefore corroborating evidence from larger samples is necessary. Second, the grammatically incorrect utterances of the grammaticality judgement may have inadvertently altered the nature of the task because storing and recalling words as part of natural utterances may be fundamentally different from doing so with ungrammatical utterances (i.e., the former reflects natural processing and thus may benefit from correctly ordered word sequences). In other words, the reading comprehension system utilizes the predictability of upcoming words based on collocations and context so using jumbled utterances may lead to a processing deficit. This can be addressed by replacing the grammaticality judgement test with an affirmative-negative judgement. In other words, the task would only be composed of natural affirmative and negative utterances.

Table 6

Standardized residual loadings for the first contrast

Loading	Item	Loading	Item
0.71	I3.3, *tsukaimashou mizuwo taisetsuni* [water let's use wisely]*	-0.67	I9.3, *nomanai sakewo amari* [much alcohol I don't drink]*
0.65	I3.2, *kouende tomodachito asobu* [I play with my friends in the park]	-0.66	I9.4, *otousanwa inuga sukida* [my father likes dogs]
0.58	I3.1, *ikenai isshouni bokuwa* [can't go with you I]*	-0.51	I10.2, *furu ashitawa amewo* [it will tomorrow rain]*
0.49	I6.3, *ekiga chikakattara benrida* [It is convenient to have the station close-by]	-0.44	I2.1, *kodomotachiga okashiwo kau* [children buy snacks]
0.49	I6.4, *dekiru konohendewa hanamiwa* [We can in this area do *hanami* (cherry-blossom viewing)]*	-0.43	I9.2, *ryokouwa denshade iku* [I am going to travel by train]
0.41	I1.2, *amakunai wasabi zenzen* [wasabi at all isn't hot]*	-0.36	I2.2, *umeboshiya nattoga kiraida* [I dislike *umeboshi* (salted plums) and *natto* (fermented beans)]
0.39	I4.3, *kaerimasu seitowa aruite* [go home the students on foot]*	-0.33	I7.1, *honwo yomisugiruto mega tsukareru* [When I read too much, my eyes get tired]
0.2	I1.1, *sono eigawa kowai* [the movie is scary]	-0.29	I9.1, *oniichanwa yakyuuwo yameru* [my brother is going to quit baseball]
0.18	I5.4, *eigono shikenwa kantanda* [the English exam is easy]	-0.28	I7.3, *utaimasu tomodachiga jouzuni* [sings my friend well]*
0.17	I6.1, *nihonwa supeinyori semai* [Japan is smaller than Spain]	-0.24	I7.2, *nerutokini denkiwo kesu* [I turn off the lights when I go to bed]
0.13	I7.5, *jibunno mochitai misega* [my own shop I want to run]*	-0.23	I10.1, *komaru tsukattara okanewo* [money I will be troubled if I spend]*
0.08	I6.2, *aitia hitowo atarashi* [a person new I want to meet]*	-0.2	I8.1, *maketa shiaiwa kinouno* [the game yesterday we lost]*
0.05	I5.1, *kotoga aru shinpaina* [there is worries me something]*	-0.11	I8.3, *oishii totemo gohanwa* [very good the food is]*
0.03	I4.2, *taberu bokuwa ringowo* [eat I apples]*	-0.11	I9.5, *koutsujikoga mainichi aru* [there are traffic accidents every day]
0.03	I5.3, *arukinikui kono kutsuwa totemo* [it's very hard on these shoes to walk]*	-0.03	I8.2, *heiyaga totemo kireida* [the room is very clean]
0	I5.2, *hashiruto ashiga itai* [it hurts when I run]	-0.02	I4.1, *kenkounotameni undousuru* [I exercise to stay healthy]

Note. *Translation written in incorrect English word order to reflect the ungrammatical Japanese sentences.

Third, the scoring system may have created interdependence among the items, inflating the reliability coefficients. To address this issue, future investigations should include a polytomous analysis of the superitems (sets treated as items) following the analysis of the individual items. In addition, future scoring systems should account for the processing component of the task, and perhaps, the word recall interval (i.e., the time between the recall of one word and the next). Fourth, this study did not account for the abstract or concrete nature of the target words. In future investigations, the number of abstract words should be controlled for as they tend to be more difficult to recall. Similarly, the target words should be selected from frequency lists because if words with overly different frequencies are mixed together in the same task, it is difficult to know if it is WM or word recognition what is causing the recall errors. Finally, future research should also examine the impact of utterance complexity on performance.

Conclusion

Despite the popularity of listening span tasks since Daneman and Carpenter's (1980) pioneering research, no study has attempted to revise and/or collect evidence to support their construct validity. This study provides initial psychometric validity evidence for a new listening span task that was constructed to address the shortcomings of length of utterance and knowledge bias identified in previous tests. The Rasch model appeared to be a suitable approach to investigate the functioning and validity of listening span tasks and, perhaps, WM measures in general. It is this author's hope that researchers employ this listening span task to further improve the assessment of WM capacity.

Notes

[1] A mora is defined as a minimal unit of sound of metrical time in the Japanese phonological system.

Acknowledgements

Many thanks to my advisor, Prof. James Sick, for his methodological guidance. I am also grateful to the reviewers and editor for their constructive feedback. Any remaining errors are my own. I would also like to express my gratitude to Andrew Wright, Clint Denison, and Sachiko Aoki for their assistance, and to all the teachers and students who made this research possible.

References

Baddeley, A. D., Kopelman, M. D., & Wilson, A. B. (2002). *The handbook of memory disorders* (2nd. Ed). Wiley.

Bazan, B. (2020). A Rasch-validation study of a novel speaking span task. *Shiken, 24*(1), 1–21. Retrieved from http://teval.jalt.org/node/95

Bond, T. G., & Fox, C. M. (2007). *Applying the Rasch model: Fundamental measurement in the human sciences* (2nd. ed.). Erlbaum.

Boone, W. J., & Noltemeyer, A. (2017). Rasch analysis: A primer for school psychology researchers and practitioners, *Cogent Education*, *4*(1),1–13. https://doi.org/10.1080/2331186X.2017.1416898

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

Conway, A. R. A., Kane, M. J., Buntig M. F., Hambrick, D. Z., Wilhelm, O., & Engle, R. W. (2005). Working memory span tasks: A methodological review and user's guide. *Psychonomic Bulletin and Review*, *12*(5), 769–786. https://doi.org/10.3758/BF03196772

Cowan, N. (1999). An embedded-processes model of working memory. In A. Miyake & P. Shah (Eds.), *Models of working memory: Mechanisms of active maintenance and executive control* (pp. 62–102). Cambridge University Press. https://doi.org/10.1017/CBO9781139174909

Daneman, M., & Carpenter, P. A. (1980). Individual differences in working memory and reading. *Journal of Verbal Learning and Verbal Behavior*, *19*(4), 450–466. https://doi.org/10.1016/S0022-5371(80)90312-6

Daneman, M., & Merikle, P. M. (1996). Working memory and language comprehension: A meta-analysis. *Psychonomic Bulletin & Review*, *3*, 422–433. https://doi.org/10.3758/BF03214546

Duncan, P. W., Bode, R. K., Lai, S. M., & Perera, S. (2003). Rasch analysis of a new stroke-specific outcome scale: The stroke impact scale. *Archives of Physical Medicine and Rehabilitation, 84*, 950–963. https://doi.org/10.1016/S0003-9993(03)00035-2

Friedman, N., & Miyake, A. (2005). Comparison of four scoring methods for the reading span test. *Behavior Research Methods*, *37*(4), 581–590. https://doi.org/10.3758/bf03192728

Gathercole, S. E., & Baddeley, A. D. (1993). *Essays in cognitive psychology. Working memory and language.* Erlbaum.

Ivanova, M. V., & Hallowell, B. (2014). A new modified listening span task to enhance validity of working memory assessment for people with and without aphasia. *Journal of Communication Disorders*, *52*, 78–98. https://doi.org/10.1016/j.jcomdis.2014.06.001

Komori, M. (2016). Effects of working memory capacity on metacognitive monitoring: A study of groups differences using a listening span test. *Frontiers in Psychology*, *7*, 1–9. https://doi.org.10.3389.2016.00285

Linacre, J. M. (2002). What do infit, outfit, mean-square, and standardized mean? *Rasch Measurement Transactions, 16*(2), 878.

Linacre, J. M. (2007). *A user's guide to WINSTEPS: Rasch-model computer program.* MESA.

Linacre, J. M. (2013). Reliability, separation, and strata: Percentage of sample in each level. *Rasch Measurement Transactions*, *26*(4), 1399. https://www.rasch.org/rmt/rmt264g.htm

Linacre, J. M. (2018). Dimensionality: Contrasts and variances. www.winsteps.com/winman/webpage.htm

Linacre, J. M. (2018). Winsteps (Version 4.3.1) [Computer Software]. Winsteps.com.

Linck, J. A., Osthus, P., Koeth, J. T., & Bunting, M. F. (2014). Working memory and second language comprehension and

production: A meta-analysis. *Psychonomic Bulletin & Review, 21,* 861–883. https://doi.org/10.3758/s13423-013-0565-2

Miyake, A., Friedman, N. P., Emerson, M. J., Witzki, A. H., & Howerter, A. (2000). The unity and diversity of executive functions and their contribution to complex frontal lobe tasks: A latent variable analysis. *Cognitive Psychology*, *41*, 49–100. https://doi.org/10.1006/cogp.1999.0734

Miyake, A. and Shah, P. (Eds) (1999). *Models of working memory: Mechanisms of active maintenance and executive control*. Cambridge University Press.

Osaka, M., Osaka, N., Kondo, H., Morishita, M., Fukuyama, H., Aso, T., & Shibasaki, H. (2003). The neural basis of individual differences in working memory capacity: An fMRI study. *Neuroimage*, *18*, 789–797. https://doi.org/10.1016/S1053-8119(02)00032-0

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests.* University of Chicago.

Reckase, M. D. (1979). Unifactor latent trait models applied to multifactor tests: Results and implications. *Journal of Educational Statistics*, *4*, 207–230. https://doi.org/10.2307/1164671

Towse, J. N., Hitch, G. J., & Hutton, U. (2000). On the interpretation of working memory span in adults. *Memory and Cognition*, *28*, 341–348. https://doi.org/10.3758/BF03198549

Turner, M. L., & Engle, R. W. (1989). Is working memory capacity task dependent? *Journal of Memory and Language*, *28*, 127–154. https://doi.org/10.1016/0749-596X(89)90040-5

Waters G. S., & Caplan, D. (2003). Verbal working memory and on-line syntactic processing: Evidence from self-paced listening. *Quarterly Journal of Experimental Psychology*, *57*, 129–163. https://doi.org/10.1080/02724980343000170

Appendix A

Listening Span Task

Set	Item	Grammaticality		Sentence
Set 1	I1.1	✓	Japanese	その映画は怖い
			Romanized Version	*sono eigawa kowai*
			English Translation	the movie is scary
	I1.2	×	Japanese	甘くないワサビは全然
			Romanized Version	*amakunai wasabi zenzen*
			English Translation	wasabi at all isn't hot*
Set	Item	Grammaticality		Sentence
Set 2	I2.1	✓	Japanese	子供たちがお菓子を買う
			Romanized Version	*kodomotachiga okashiwo kau*
			English Translation	children buy snacks
	I2.2	✓	Japanese	梅干しや納豆が嫌いだ
			Romanized Version	*umeboshiya nattoga kiraida*
			English Translation	I dislike *umeboshi* (salted plums) and *natto* (fermented beans)
Set	Item	Grammaticality		Sentence
Set 3	I3.1	×	Japanese	行けない一緒に僕は
			Romanized Version	*ikenai isshouni bokuwa*
			English Translation	can't go with you I*
	I3.2	✓	Japanese	公園で友達と遊ぶ
			Romanized Version	*kouende tomodachito asobu*
			English Translation	I play with my friends in the park
	I3.3	×	Japanese	使いましょう水は大切に
			Romanized Version	*tsukaimashou mizuwo taisetsuni*
			English Translation	water let's use wisely*
Set	Item	Grammaticality		Sentence
Set 4	I4.1	✓	Japanese	健康のために運動する
			Romanized Version	*kenkounotameni undousuru*
			English Translation	I exercise to stay healthy
	I4.2	✓	Japanese	食べる僕はリンゴを
			Romanized Version	*taberu bokuwa ringowo*
			English Translation	eat I apples
	I4.3	×	Japanese	帰ります生徒はあるいて
			Romanized Version	*kaerimasu seitowa aruite*
			English Translation	go home the students on foot*
Set	Item	Grammaticality		Sentence
Set 5	I5.1	×	Japanese	ことがある心配な
			Romanized Version	*kotoga aru shinpaina*
			English Translation	there is worries me something*
	I5.2	✓	Japanese	走ると足が痛いだ
			Romanized Version	*hashiruto ashiga itai*
			English Translation	it hurts when I run
	I5.3	×	Japanese	歩きにくいこの靴はとても
			Romanized Version	*arukinikui konokutsuwa totemo*
			English Translation	it's very hard on these shoes to walk*
	I5.4	✓	Japanese	英語の試験は簡単だ
			Romanized Version	*eigono shikenwa kantanda*
			English Translation	The English exam is easy
Set	Item	Grammaticality		Sentence
	I6.1	✓	Japanese	日本はスペインより狭い
			Romanized Version	*nihonwa supeinyori semai*
			English Translation	Japan is smaller than Spain
			Japanese	会いたい人を新しい

Set 6	I6.2	×	Romanized Version	*aitai hitowo atarashii*
			English Translation	a person new I want to meet*
			Japanese	駅が近かったら、便利だ
	I6.3	✓	Romanized Version	*ekiga chikakattara, benrida*
			English Translation	It is convenient to have the station close-by
			Japanese	出来るこの辺では花見は
	I6.4	×	Romanized Version	*dekiru konohendewa hanamiwa*
			English Translation	We can in this area do *hanami* (cherry-blossom viewing)*

Set	Item	Grammaticality		Sentence
			Japanese	本を読みすぎると目が疲れる
	I7.1	✓	Romanized Version	*honwo yomisugiruto mega tsukareru*
			English Translation	When I read too much, my eyes get tired
			Japanese	寝る時に電気を消す
	I7.2	✓	Romanized Version	*nerutokini denkiwo kesu*
			English Translation	I turn off the lights when I go to bed*
Set			Japanese	歌います友達は上手に
7	I7.3	×	Romanized Version	*utaimasu tomodachiwa jouzuni*
			English Translation	sings my friend well
			Japanese	僕はお金が欲しいだ
	I7.4	✓	Romanized Version	*bokuwa okanega hoshii*
			English Translation	I want money
			Japanese	自分の持ちたい店が
	I7.5	✓	Romanized Version	*jibunno mochitai misega*
			English Translation	my own shop I want to run

Set	Item	Grammaticality		Sentence
			Japanese	負けたしあいは昨日の
	I8.1	×	Romanized Version	*maketa shiai kinouno*
Set			English Translation	the game yesterday we lost*
8			Japanese	部屋がとてもきれいだ
	I8.2	✓	Romanized Version	*heiyaga totemo kireida*
			English Translation	the room is very clean
			Japanese	美味しいとてもご飯は
	I8.3	×	Romanized Version	*oishii totemo gohanwa*
			English Translation	very good the food is*
			Japanese	この飲みにくい薬は
	I8.4	×	Romanized Version	*kononominikui kusuriwa*
			English Translation	hard to swallow the medicine is*
			Japanese	間に合う次の電車に
	I8.5	×	Romanized Version	*maniau tsugino denshani*
			English Translation	in time we are for the next train*

Set	Item	Grammaticality		Sentence
			Japanese	お兄ちゃんは野球を辞める
	I9.1	✓	Romanized Version	*oniichanwa yakyuuwo yameru*
			English Translation	my brother is going to quit baseball
			Japanese	旅行は電車で行く
	I9.2	✓	Romanized Version	*ryokouwa denshade iku*
			English Translation	I am going to travel by train
			Japanese	飲まない酒をあまり
	I9.3	×	Romanized Version	*nomanai sakewo amari*
			English Translation	much alcohol I don't drink
Set			Japanese	お父さんは犬が好きだ
9	I9.4	✓	Romanized Version	*otousanwa inuga sukida*
			English Translation	my father likes dogs
			Japanese	交通事故が毎日ある
	I9.5	×	Romanized Version	*koutsujikoga mainichi aru*
			English Translation	there are traffic accidents every day*
			Japanese	授業はもう始まった

	I9.6	✓	Romanized Version	*jugyouwa mou hajimatta*
			English Translation	the class has already started
Set	Item	Grammaticality		Sentence
Set 10	I10.1	×	Japanese	困る使ったらお金を
			Romanized Version	*komaru tsukattara okanewo*
			English Translation	money I will be troubled if I spend*
	I10.2	×	Japanese	降る明日は雨が
			Romanized Version	*furu ashitawa amega*
			English Translation	it will tomorrow rain*
	I10.3	×	Japanese	この問題は難しいです
			Romanized Version	*konomondaiwa muzukashii desu*
			English Translation	this question is difficult
	I10.4	✓	Japanese	山にヤギがいた
			Romanized Version	*yamani yagiga ita*
			English Translation	there were goats in the mountains
	I10.5	×	Japanese	僕は帰ると思う
			Romanized Version	*bokuwa kaeruto omou*
			English Translation	I think I will go home*
	I10.6	×	Japanese	くれた送って友達が
			Romanized Version	*kureta okutte tomodachiga*
			English Translation	home took me my friend*

Note. *Translation written in incorrect English word order to reflect the ungrammatical Japanese sentences.

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.