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

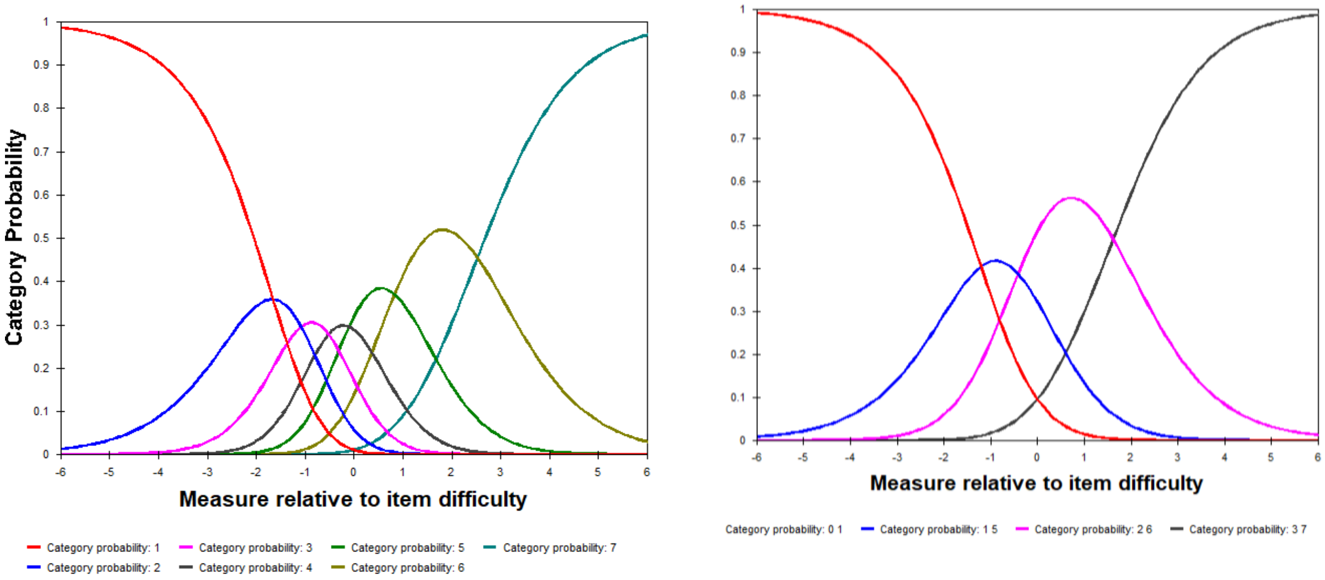


Table 1

Summary of scale category structure for phonological similarity

					%		*M*		Infit		Outfit		Andrich		Category	
Rating	Score		Count		Observed		(Logits)		*MS*		*MS*		Threshold		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

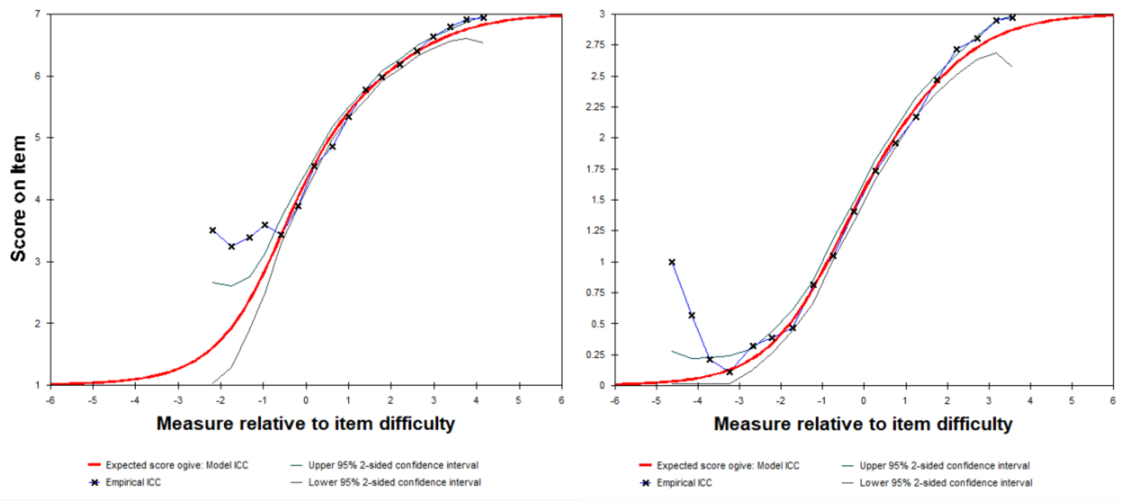


Figure 3

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

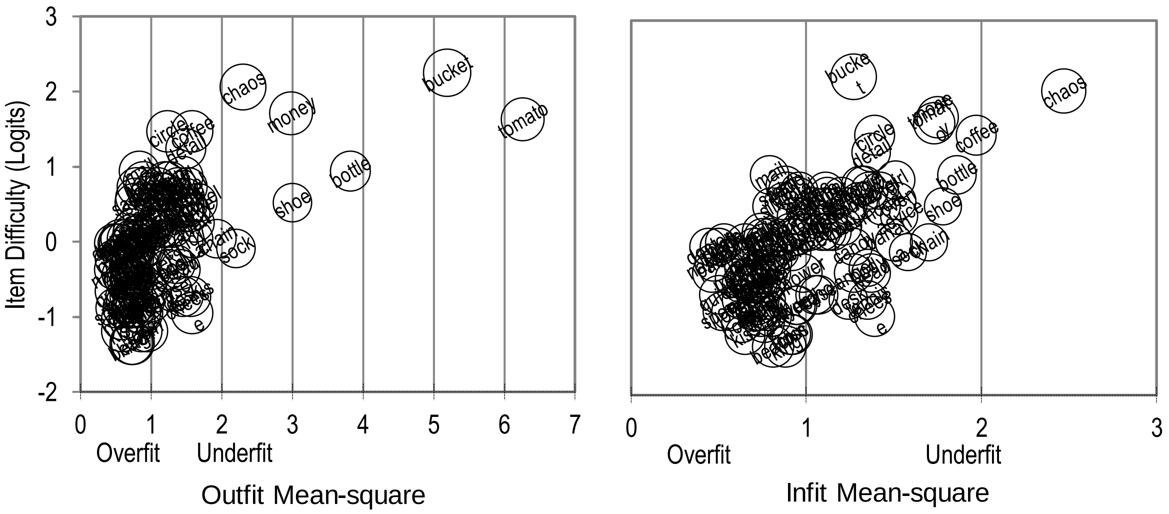
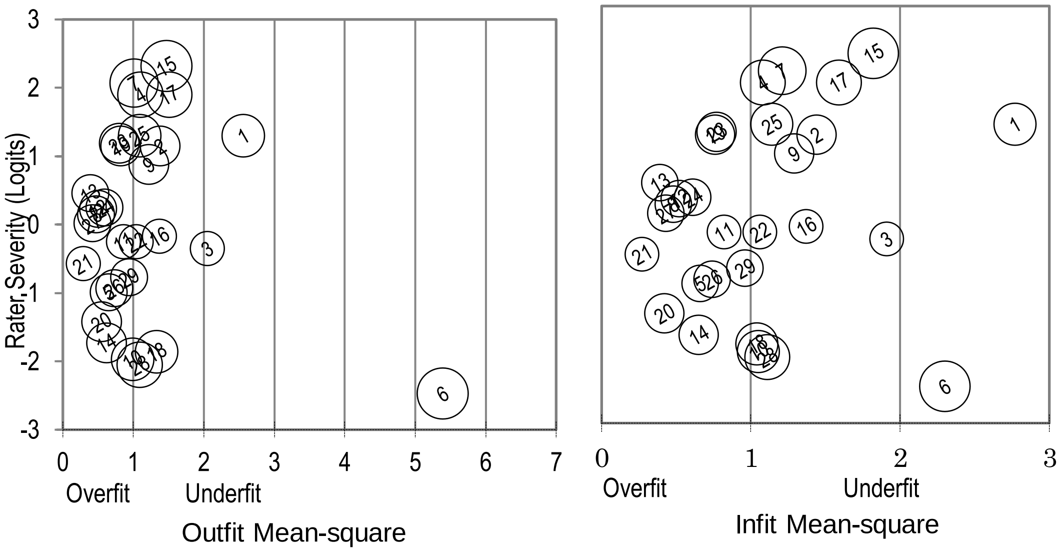


Figure 4 shows the person pathway maps, with Raters 1, 3, and 6 having outfit mean-square values exceeding 2.00 and high infit values. Raters 17 and 15 also showed high infit values, but these raters were extremely lenient, meaning they judged nearly all words to be highly similar. Rater 6 was extremely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

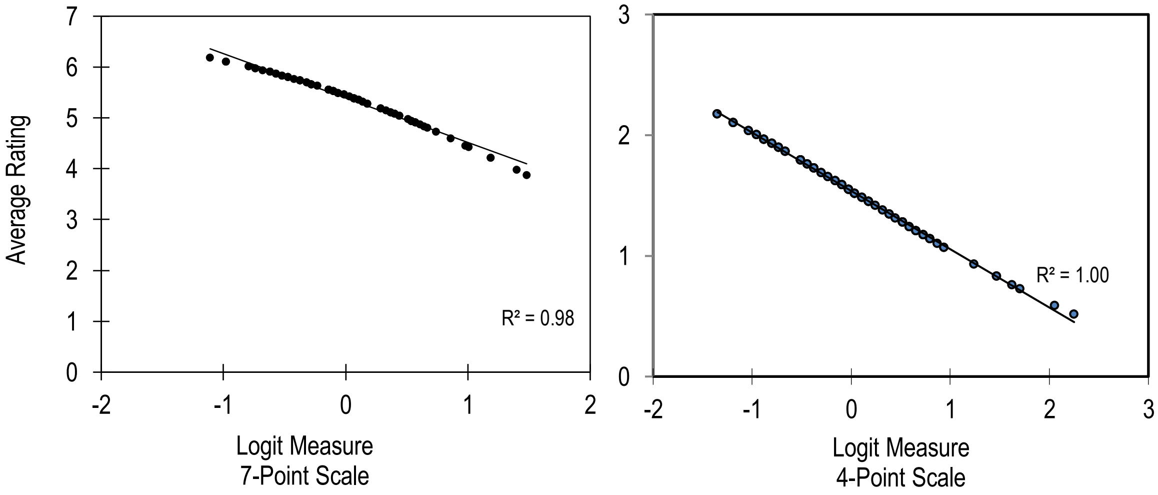


To further investigate rater idiosyncrasy, an analysis was conducted of the most unexpected responses, that is, those responses with the largest standardized residuals from the rescored scale. Of the 20 most unexpected responses, nine came from Rater 6, further highlighting her as behaving extremely 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

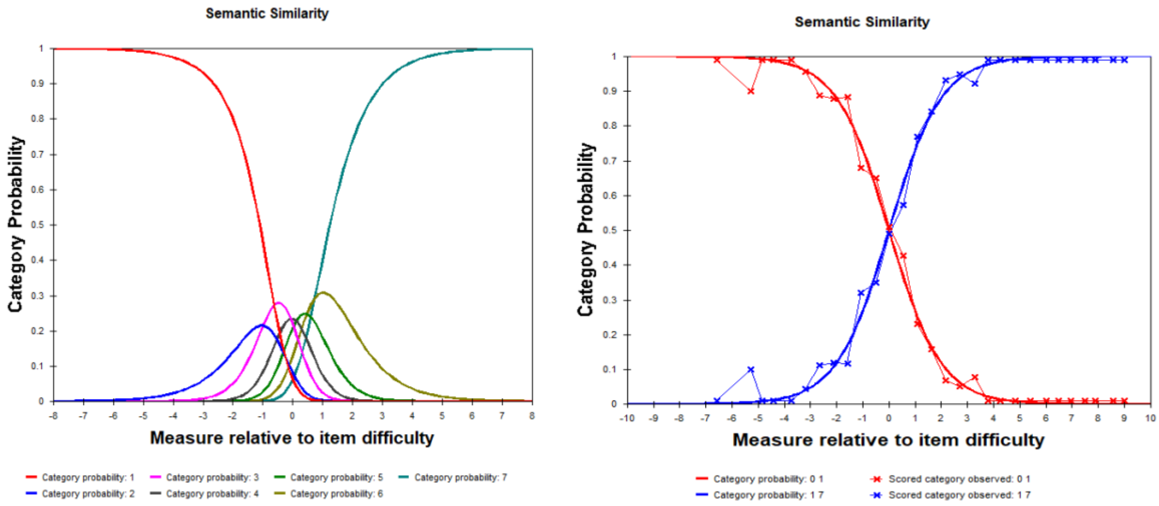


Table 5

Summary of scale category structure for semantic similarity

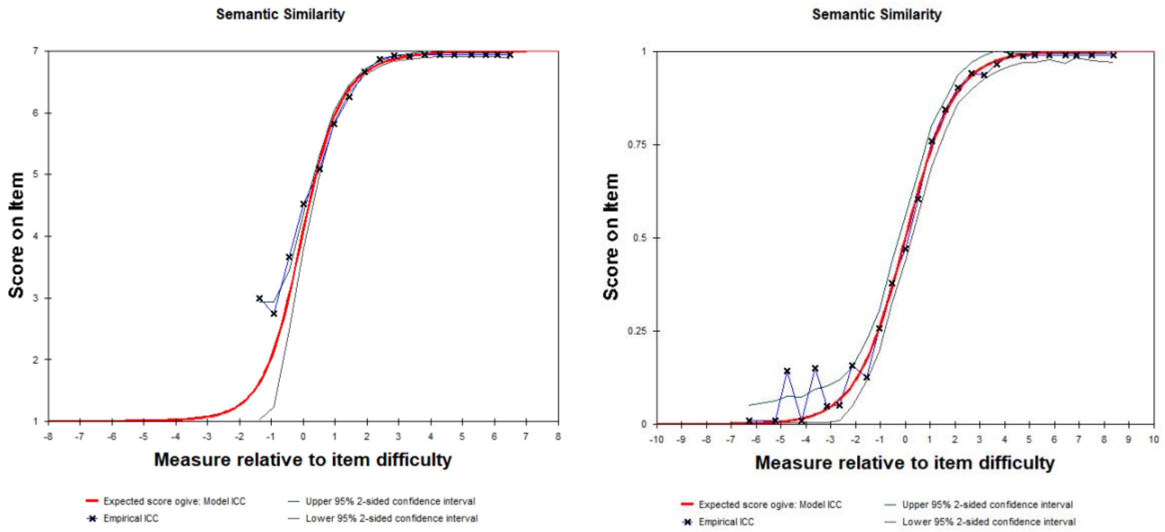
Rating	Score		Count		% Observed		*M* (Logits)		Infit *MS*		Outfit *MS*		Andrich Threshold		Category Measure	
	7-Pt	2-Pt	7-Pt	2-Pt	7-Pt	2-Pt	7-Pt	2-Pt	7-Pt	2-Pt	7-Pt	2-Pt	7-Pt	2-Pt	7-Pt	2-Pt
			23		1											n.a.
1	1	0					0.06		1.58		2.27		NONE	n.a.	(-2.05)	
2	2	0	26		1		0.28		1.45		1.76		-0.25	n.a.	-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

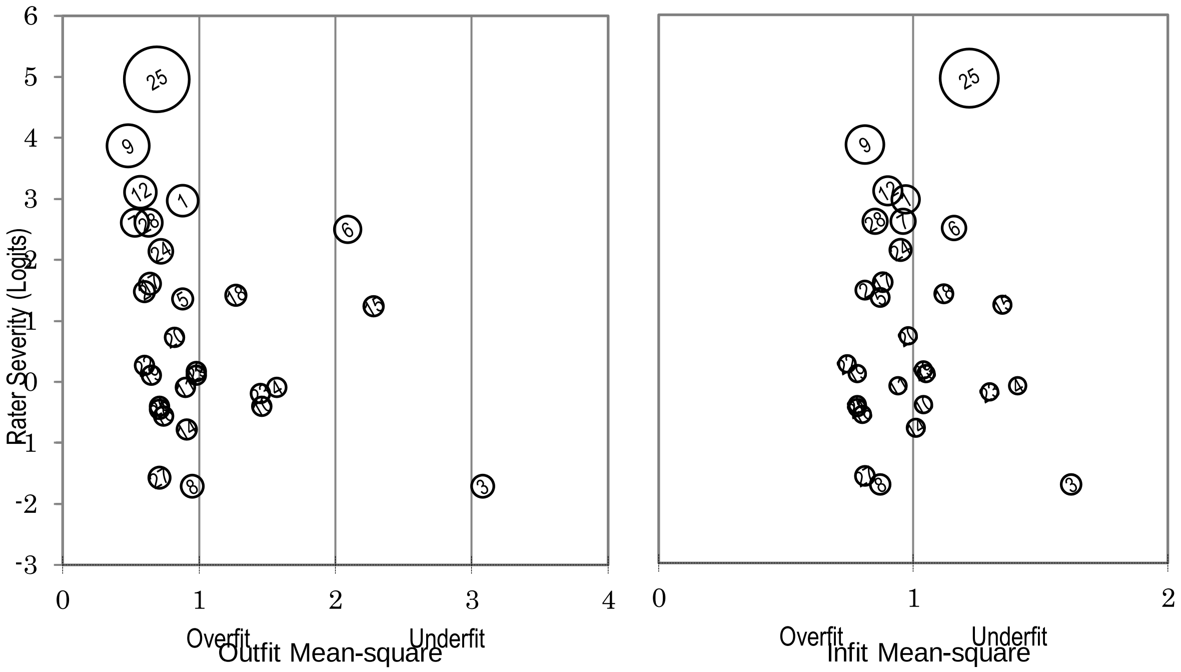


Figure 9

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

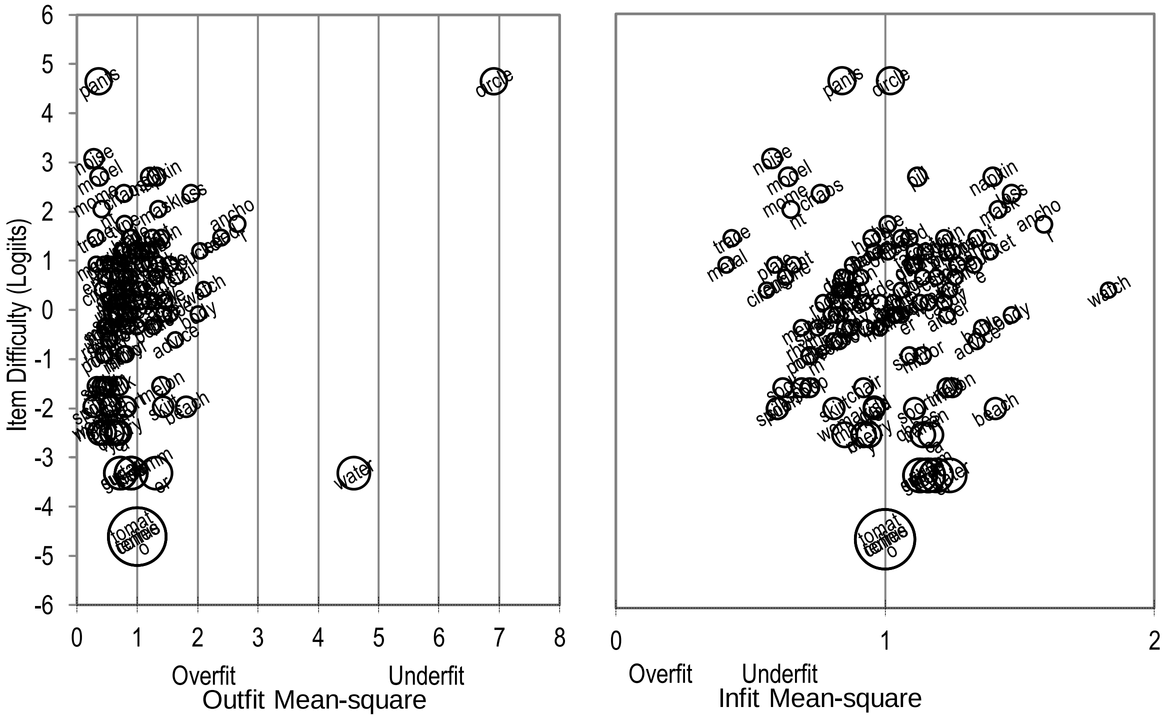
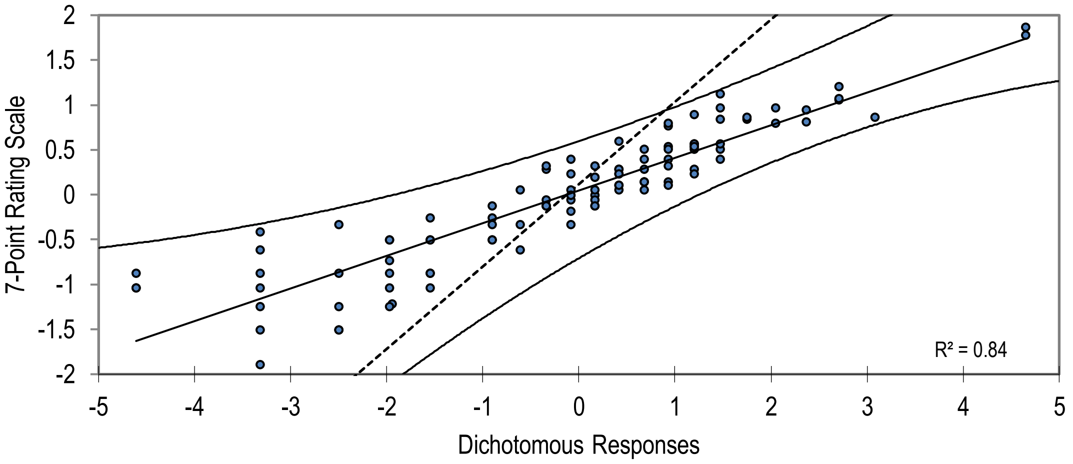


Figure 10 compares item difficulty (i.e., semantic similarity rating) from the 7-point rating scale and dichotomous ratings. No items fall outside the 95% confidence intervals, with the 84% shared variance reflecting a raw correlation of .93, rising to 1.00 after disattenuation for reliability. Figure 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

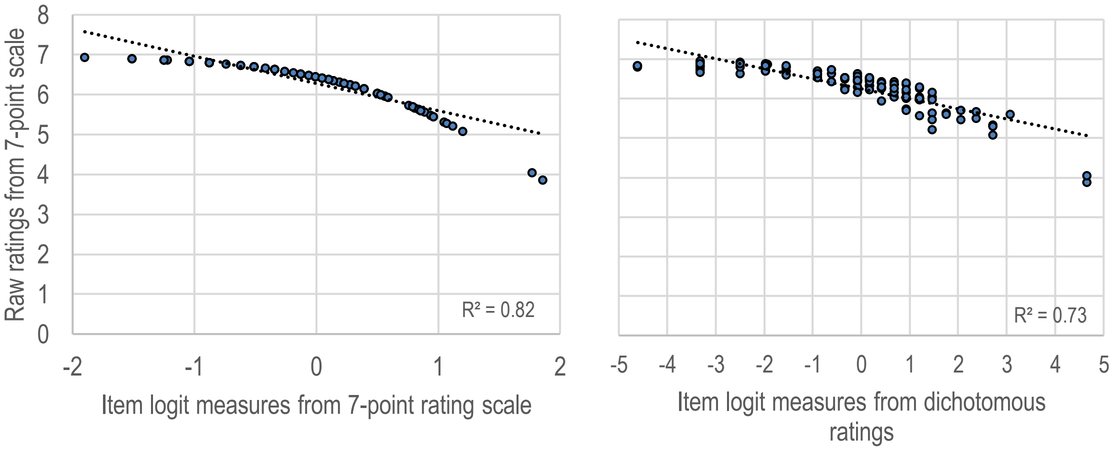
Comparison of semantic similarity ratings from 7-point rating scale and rescaled dichotomous data



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 /uɪndo/ and *wing*-ウイング /uɪngu/, 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.