A response to Holster and Lake regarding guessing and the Rasch model

Jeffrey Stewart  
*Kyushu Sangyo University*

Stuart McLean  
*Osaka Jogakuin University*

Brandon Kramer  
*Osaka Jogakuin University*

**Abstract**

Stewart (2014) questioned vocabulary size estimation methods proposed by Beglar and Nation for the VST, further arguing Rasch mean square (MSQ) fit statistics cannot determine the proportion of random guesses contained in the average learner’s raw score, as the average value will be near 1 by design. He illustrated this by demonstrating this is true even of entirely random data. Holster and Lake (2016) appear to misinterpret this as a claim that Rasch analyses cannot distinguish random data from real responses. To test this, they compare real data to random and note that, predictably, the statistic easily distinguishes the two, and that reliability for random data is near zero. However, while certainly true, this fact is not relevant to Stewart’s argument that multiple-choice options inflate the test’s size estimates, or that MSQ statistics cannot be used to detect this. We further illustrate this by showing real data retains average MSQ values near 1 even when unknown items skipped by learners are imputed with random guesses. Furthermore, the imputed data does not exhibit “problematic guessing” under Holster and Lake’s own criteria, despite size inflation under Beglar and Nation’s suggested scoring. We conclude by discussing uses of the 3PL model.

**Keywords:** Vocabulary Testing, The Rasch Model, Item Response Theory
A Response to Holster and Lake Regarding Guessing and the Rasch Model

A recent LAQ commentary by Holster and Lake (2016) on guessing and the Rasch model extensively cites and discusses an earlier commentary by Stewart (2014) on inflated estimates of vocabulary size made by the Vocabulary Size Test (VST) (Nation & Beglar, 2007). Unfortunately, Holster and Lake appear to have misunderstood Stewart’s observations regarding the Rasch model and suggestions related to analysis of the VST under the 3-parameter logistic (3PL) model (Birnbaum, 1968). This response will clarify Stewart’s positions regarding suggested uses of these item response models and the nature of Holster and Lake’s misunderstanding of them. We will begin with a summary of the original papers and major points of confusion, detail an empirical study using real data which re-illustrates Stewart’s position on uses of MSQ fits statistics as they relate to size estimation under the VST, and briefly discuss recommended (and not recommended) uses of the 3PL model as it relates to research on the VST.

Stewart (2014) expressed a concern about a common method of estimating learner vocabulary size using the VST (Nation & Beglar, 2007), which was endorsed in both the original publication on the test, and in separate subsequent papers by its coauthors (Beglar, 2010; Nation, 2012). As Beglar explained in his later Rasch-based validation of the test, "The most straightforward way to interpret the VST results is to view each item as representing 100 word families (assuming that 10 items are used at each 1000-word frequency level). A test-taker’s raw score would be multiplied by 100 to obtain their total vocabulary size" (Beglar, 2010, p. 114). For clarity, we will hereafter refer to this straightforward suggestion as the simple multiplication rule: A single correctly-answered multiple choice question is assumed to equal knowledge of the tested word family. Therefore, using a simple polling method, the proportion of word families known within a given 1000 word-family level is estimated by examining how many of the ten multiple-choice questions testing the level were answered correctly, and multiplying this number by 100.

Stewart noted the danger of this assumption: in addition to the question of whether or not answering such items can be regarded as evidence a learner sufficiently “knows” a word, even when learners encounter words they are entirely unfamiliar with, they may still correctly answer approximately 25% of these 4-option multiple-choice items by chance. This would, in turn, inflate size estimates made using the simple multiplication rule. He further demonstrated that the best-known validation of the VST by Beglar (2010), which relied on Rasch analysis, was not sufficient to detect the accuracy of the simple multiplication rule in this respect; while Rasch analysis has many useful applications, determining the precise proportion of multiple-choice questions that were answered by chance is not one of them, as this information is typically of little relevance in most testing contexts, provided the test’s internal reliability remains sufficiently high. Rather, analyses used in Rasch measurement are of use in assessing the overall item and person reliability of the test, determining if the tested construct can be considered to be essentially unidimensional, and establishing if the test has adequate (and therefore useful) fit to
the Rasch model. In his conclusion, for research purposes Stewart recommended conducting an analysis of the VST under the 3PL model, which, unlike the Rasch model, accounts for above-zero likelihoods that test takers of very low ability can correctly guess multiple-choice items.

Subsequent empirical research appears to confirm Stewart's thesis that the simple multiplication rule results in inflated counts of learner vocabulary size. In a recent study, we determined that the 3PL model has superior fit to the first eight levels of the 14,000 word family VST (Mclean, Kramer, & Stewart, 2015). Other researchers have explored the effect of guessing on VST scores by interviewing learners after their completion of the test; Gyllstad, Vilkaitė and Schmitt (2015) reported a substantial inflation of size estimates due to guessing. Perhaps due to such findings, there appears to be a growing consensus within the field of vocabulary testing that the simple multiplication rule be abandoned, and as Nation recently stated, “I am being convinced that the benefits from using the whole test (truly known low frequency words) are outweighed by the disadvantages (guessing)” (I.S.P Nation, personal communication, December 2015).

In their recent response to Stewart's paper, Holster and Lake agree with Stewart's central thesis that Beglar and Nation’s suggested use of raw scores for estimates of vocabulary size is problematic. In their study, they replace Beglar and Nation’s suggested simple multiplication rule with a simple correction for guessing formula which accounts for the proportion of incorrect attempted answers. Adjustment formulas for multiple and selected-response vocabulary tests are an area of increasing interest in vocabulary testing (e.g., Stewart & White, 2011; Stubbe & Stewart, 2012). Although this practice was explicitly rejected by Nation (2012, p. 6) and our preferred resolution would be a new version of the VST using an item format for which minimal guessing can be assumed, we do believe that use of such an adjustment would mark an improvement to the test’s size estimates.

However, as proponents of Rasch measurement, Holster and Lake take strong exception to Stewart's argument that Rasch analysis is not useful in detecting “problematic” guessing in multiple-choice tests. In particular, they object to a demonstration by Stewart that mean square (MSQ) fit statistics used in Rasch analysis to detect problematic guessing have average values near one by definition, regardless of the quality of the underlying data. Stewart had demonstrated this by showing that this is the case even for a data set comprised entirely of random responses. In response, Holster and Lake concurrently analyze both real responses and randomly generated responses and demonstrate that, predictably, the statistic can easily distinguish between the two.

They further point out that under Rasch analysis, random responses will obviously have person and item reliability statistics near zero.

Holster and Lake’s observations are straightforward and universally accepted. As frequent users of the Rasch model in our own work and research (e.g., McLean, Kramer, & Beglar, 2015; Stewart, Batty, & Bovee, 2012), we appreciate its many applications, and understand their desire to defend the practice of Rasch measurement from what they mistakenly perceive as criticism. However, while the observations made in their study are certainly true,
these points are not challenged in Stewart (2014), and are not relevant to Stewart’s argument that multiple-choice options inflate size estimates on the VST and that MSQ fit statistics cannot be used to detect this.

Holster and Lake’s confusion appears to stem from a misunderstanding regarding how “problematic” guessing is defined. In Stewart's manuscript, guessing on the VST is “problematic” in the sense that its multiple-choice format will cause the simple multiplication rule to return inflated size estimates derived from unadjusted raw scores, regardless of the tests’ high internal reliability or its adequate (and therefore, otherwise still useful) fit to the Rasch model (Stewart, 2014, p. 276). In contrast to this, in Holster and Lake's manuscript, “problematic” guessing appears to refer to guessing behaviors which could threaten to “change the relative positions of persons and items on the latent trait beyond the limits of measurement error” (2016, p. 124). As raw scores are a sufficient statistic for person ability estimates under the Rasch model (De Ayala, 2009, p. 25), and therefore the two can be expected to have high correlations regardless of the quality of the analyzed data, it is uncontroversial to state that guessing on the VST need not be "problematic" in this respect. However, no claim to the contrary is made in Stewart (2014), and this fact is of little relevance to his argument. We will further illustrate this using real data.

A further point of contention is Stewart's recommendation that a 3PL analysis be conducted on the VST for the purpose of determining minimum probabilities of correct guesses on difficult VST items by low-level learners. Holster and Lake appear to interpret this suggestion as a wholesale endorsement of the 3PL model not only for a single analysis of VST item parameters for research purposes, but also for scoring by all educators and researchers for all future uses of the test, despite the numerous issues and challenges such a proposition would pose. As they argue, scoring the VST under the 3PL is problematic on theoretical and practical grounds. However, their arguments on this matter, too, are not relevant to Stewart’s paper. Stewart recommends only a single analysis under the 3PL for research purposes, emphasizing the model's analytical properties rather than its disputed measurement properties (Stewart, 2014, p. 280), and as we explained in an empirical analysis of the VST using the 3PL, we do not recommend scoring the VST under the model (Mclean, Kramer, & Stewart, 2015, pp. 26–27). The final section of this response will further explain this distinction.

**Use of Mean Square Fit Statistics to detect “problematic” guessing**

As noted above, Stewart argued that MSQ fit statistics cannot be used to determine the proportion of the average (i.e., non-outlier) learner’s raw score that can be attributed to guessing. He illustrated this by demonstrating that even a data set comprised entirely of simulated random guesses returns average MSQ fit statistic values of 1, under the reasoning that if values remain near 1 under even such extreme circumstances, they will surely also remain near 1 in cases where real learners are instructed to guess the answers to vocabulary items well above their proficiency level that they do not know, in addition to the higher-frequency vocabulary items
that they do know. Granted, this scenario was far from what one would see in a real testing context. But then, so is the scenario presented in Holster and Lake’s experiment, in which half the data was real, and half entirely comprised of random guesses. In practice, it is rare for more than a small percentage of test takers to guess every item on a test entirely at random. Rather, learners are more likely to answer questions they believe they know the answer to in good faith, and attempt to guess the answers to questions they do not.

We will re-illustrate Stewart’s argument under this more realistic scenario. To do this, we used a real VST data set of Japanese university students ($N = 354$) attempting the first eight levels of the 14,000-word family version of the VST. Learners were instructed to skip items testing words they did not believe they knew the answers to. This resulted in a dataset with a total of 12,099 correctly answered items, 9610 attempted incorrectly answered items, and 6611 items flagged by learners as unknown and skipped. We then imputed the skipped items with random guesses with a 0.25 probability of chance success, and, as per Holster and Lake, compared not only MSQ fit statistics between the changed and unchanged data set, but also person reliability and Rasch PCA’s of the data’s residuals using the software program Winsteps (Linacre, 2011).

### Table 1.

<table>
<thead>
<tr>
<th>Effects of imputing skipped, unknown VST items with random guesses</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
</tr>
<tr>
<td>Mean Total Score</td>
</tr>
<tr>
<td>Standard Deviation</td>
</tr>
<tr>
<td>Person Reliability</td>
</tr>
<tr>
<td>Mean Infit MSQ</td>
</tr>
<tr>
<td>Infit SD</td>
</tr>
<tr>
<td>Mean Outfit MSQ</td>
</tr>
<tr>
<td>Outfit SD</td>
</tr>
<tr>
<td>Misfitting Persons (Outfit &gt;1.5)</td>
</tr>
<tr>
<td>Misfitting Persons (Infit &gt;1.5)</td>
</tr>
<tr>
<td>PCA total variance predicted by Rasch Model</td>
</tr>
<tr>
<td>PCA total variance explained</td>
</tr>
<tr>
<td>1st contrast variance predicted by Rasch Model</td>
</tr>
<tr>
<td>1st contrast variance explained</td>
</tr>
</tbody>
</table>

In accordance with Stewart’s original illustration, average MSQ values remain very close to 1 after adding random guesses. As an aside, it is perhaps worth noting that considerably fewer learners are flagged as misfitting the Rasch model by this statistic after random guesses are added; the more uniform noise that is added to the data, the narrower the standard deviations of MSQ values become, and therefore the fewer learners pass the relatively conservative 1.5 threshold for flagging misfit. We suspect it is this property of MSQ fit statistics that caught Holster and Lake’s attention upon reading Stewart’s paper and performing their own experiment.
in response, as they make several references to random data appearing to display good data-model fit under it. However, while this result may be surprising, no claim is made in Stewart’s paper that analyses used in Rasch measurement cannot discern if data is random, nor that average values of 1 imply otherwise acceptable data, and this aspect of the illustration is not relevant to his argument regarding size estimates of the VST.

Holster and Lake went on to argue that effects of random guessing can further be assessed by checking the reliability of the resultant measures. This is certainly true. However, the results show that even after imputation of guesses, person reliability remains relatively high. Test person reliability suffers after imputing unknown items with guesses; here, it falls from 0.92 to 0.88. However, in Holster and Lake’s paper it is argued that a person reliability of 0.78 indicates the data set does not include “problematic” guessing. Although learners only sat the first eight levels of the VST in this study, presumably, considerably more guessing behaviour could be added to the data set before attaining person reliability as low as Holster and Lake’s own figure.

In regards to a principal components analysis of residuals, Holster and Lake note that when analyzing the random data alone, variance explained by the secondary dimension exceeds variance explained by the Rasch measures (2016, p. 133). However, this is not the case for real data with guesses added for skipped, unknown items. It is true that as with reliability, total variance explained by the Rasch measures lowers after adding guessing. However, the percentage of variance explained by the measures remains slightly above the value predicted by the model, and the percentage of variance explained by the secondary dimension (first contrast) does not change after adding the imputed guesses. This indicates the guess-imputed test would not be considered to exhibit “problematic guessing” under Holster and Lake’s own criteria, despite the separate issue of inflation of vocabulary size estimates under Nation (2012) and Beglar’s (2010) suggested method for calculating vocabulary size using the test’s raw scores.

It should be emphasized that these results do not suggest that statistics used in Rasch measurement are not useful for their intended purposes, nor that there are necessarily major undetected problems in multiple-choice vocabulary tests that simply rank learners by general written receptive vocabulary knowledge (without working under the assumption that a correct answer is always equivalent to knowledge of the tested word). Therefore, Holster and Lake are correct in stating that Rasch analysis does not fail to detect guessing that is “problematic” in any conventional sense. Such guessing may be evenly distributed enough across the data set that it does not pose a major threat to measurement, although of course test reliability will suffer somewhat, and outlier learners that guess substantially more than average can easily be detected. However, the above example does illustrate that MSQ fit statistics cannot tell us the proportion of guesses contained in the average (i.e., non-outlier) learner’s raw score. This is not due to a failing of the statistic. Rather, it is because the statistic was not designed for this unorthodox purpose.
Analysis of the Vocabulary Size Test using the 3PL Model

As Holster and Lake note, “in [a selected response] test, the actual probability of success will not approach a lower asymptote of zero…but will be somewhat higher because random guessing is possible” (p. 132). For this reason, Stewart recommended conducting a 3PL analysis on the VST “for the purpose of examining the lower asymptotes of items used” (2014, p. 280), in order to establish lower thresholds for probabilities of correct guesses. This information can be used to more accurately determine, for example, which 1000-word levels of the test learners of given proficiency levels should sit. It is true that test scoring under the 3PL model can result in learners receiving less credit for items that the model determines they are more likely to have answered correctly by chance (Chiu & Camilli, 2012). This practice would indeed be considered problematic by proponents of Rasch measurement and perhaps many other observers, particularly if the VST was given as a high-stakes test rather than as a diagnostic or research instrument. However, analysis of items’ lower asymptotes is of value even if such scoring is not employed. As Stewart argued, although the debate between proponents of Rasch measurement and alternate item response models is ongoing, “the 3PL model offers useful analytical properties even if its measurement properties are disputed” (2014, p. 280).

Despite this qualification, Holster and Lake appear to interpret this suggested analysis of item lower asymptotes as a wholesale endorsement of not only analyzing but also scoring the VST using the 3PL model, stating, “Stewart's (2014) advocacy of 3PLIRT for guessing correction, therefore, has major implications for both large-scale high-stakes tests and classroom tests” (p. 127), and claiming Stewart “sees unexpected response patterns as something requiring correction from 3PLIRT analysis” (p. 139). In response to this perceived position, Holster and Lake extensively quote Chiu and Camilli’s (2012) criticisms of 3PL scoring. However, scoring the VST under the 3PL is not proposed at any point in Stewart’s paper. As Chiu and Camilli notes themselves, “many testing programs that use the 3PL model do not use 3PL scoring,” and in such cases, the points raised in their paper “are not a central concern” (p. 83). For more information on analytical uses of the 3PL, we direct the reader to a recent article by the authors (McLean, Kramer, & Stewart, 2015), in which we conduct Stewart’s recommended analysis on the first eight levels of the VST.

Conclusion

In conclusion, although passages in Holster and Lake’s (2016) commentary left us with the impression that their study intended to address arguments made in Stewart (2014) and that their results revealed evidence against them (e.g., “Stewart’s major claims are easily testable” (p. 124); “These results show that Stewart’s (2014) conjectures lack an empirical basis” (p. 140)), their experiment is not, in fact, directly related to claims made in that paper, and is instead more useful as part of a more general discussion of how guessing is typically viewed and handled under the Rasch model.

Holster and Lake’s positions are in alignment with well-known advocates of Rasch


measurement such as Wright (1997) and Bond and Fox (2015), who have long-held objections to use of the 3PL model. The dispute between proponents of Rasch measurement and alternate item response models has had a long and sometimes heated history (Hambleton, 1994). Perhaps Stewart’s observations were perceived as a renewed challenge to Rasch measurement in favor of alternative models, sparking this longstanding debate once more. Therefore, in closing we would like to stress that our observations regarding Rasch validations of the VST concern a rather unique score interpretation made on a specific test, and should not be interpreted as a blanket endorsement of a competing theoretical framework for psychological measurement. As practitioners of Rasch measurement ourselves, we believe that the Rasch model is a standard of measurement worth aspiring to, and that whenever possible, it is preferable to remove inadequate items from its Procrustean bed rather than compromise our measurement model to reflect those items’ weaknesses. Indeed, the light cast on tests we have analyzed under other response models has often only strengthened this resolve. However, it remains the case that the Rasch model and its associated statistics cannot answer questions they were not designed to answer.

References


