Effects of the Number of Response Categories on Rating Scales

Roundtable presented at the annual conference of the American Educational Research Association, Vancouver, British Columbia

April 15, 2012

Julia Shaftel
University of Kansas

Brooke L. Nash
Ascend Learning, LLC

Susan C. Gillmor
University of Kansas

Corresponding author:
Julia Shaftel
Center for Educational Testing and Evaluation
University of Kansas
1122 West Campus Road, Room 744
Lawrence, KS 66049
785-864-0733
jshaftel@ku.edu
Abstract

Educational and psychological scales and instruments are commonly used to measure opinions, personality traits and other phenomena in education, psychology, health, and marketing. One of the primary considerations in scale development is the choice of response options or categories. Given concerns about using ordinal data for statistical analyses that typically require interval-level assumptions, it is worth investigating how closely polytomous items can measure interval-level data. The research issue for this exploratory study was to use item response theory (IRT) to compare the effects of the number of response categories and verbal anchoring on Likert-scale items as determined by effective category use and distance between categories as well as distributions associated with each type of scale.
Theoretical Framework

Level of Measurement

Many statistical procedures used in educational and psychological research require the assumption of a normally distributed dependent variable, which presumes at least interval-level data (Harwell & Gatti, 2001). However, most psychological instruments reflect an ordinal scale, which is characterized by unequal relative distances between categories with respect to the construct of interest. When the measured attribute is assumed to have an underlying continuous normal distribution and a large number of response categories is available, interval-level assumptions are often made (Howe, 2006; Tabachnik & Fidell, 2001). The practice of employing statistical techniques with ordinal-level data appears to be common in educational and psychological research. Issues such as control of Type I error and subsequent interpretation problems can arise when ordinal data are employed in statistical analyses that require interval-scale variables (Harwell & Gatti, 2001). Even the use of coefficient alpha is suspect with ordinal data as it estimates the lower bound of internal consistency using Pearson product-moment correlations among the items rather than the more appropriate polychoric correlations for ordinal data (Howe, 2006).

Scale Development Issues

Some researchers contend that multiple-category response scales may provide advantages over dichotomous scales with respect to reliability and validity (Bandalos & Enders, 1996). Comrey (1988) argued that multiple-category response scales provide more reliable and stable results; Preston and Coleman (1999) echoed these findings. Nunnally (1978) found reliability to be a monotonically increasing function of the number response categories in a given scale, though reliability levels off with little gain after 11 categories. On the other hand, Matell and
Jacoby (1971) found no differences in reliability for scale formats that ranged from 2 to 19 response categories, though they did find that selection of a neutral midpoint decreased with increasing numbers of response options. Others have also concluded that reliability coefficients remain unchanged regardless of the number of response categories (Aiken, 1983; Bendig, 1954), particularly after four or five scale points (Lozano et al., 2008; Jenkins & Taber, 1977). Andrews (1984) stated that the number of response categories was one of the most important factors influencing the construct validity of a scale, though the benefits acquired from adding scale points diminished after five points.

Modern scale formats have also been evaluated to determine their potential benefits and disadvantages. For instance, visual analog scales allow for any range of responses along a continuum typically defined at the endpoints. These scales are difficult and time consuming to score by hand and are most useful when the construct of interest is simplistic and singular (Clark & Watson, 1995). However, Russell and Bobko (1992) found that data gathered using a line segment scale produced larger effect sizes in regression analyses compared to a five-point Likert scale, suggesting that greater measurement precision was obtained using the former.

Labeling scale categories with numeric values, as opposed to label descriptions, may provide data that is more than ordinal even if not quite interval (Kiess & Bloomquist, 1985). Schwarz et al. (1991) investigated how numeric values assigned to category options impacted respondents’ results and concluded that respondents do in fact use the values to interpret the meaning of the labels. Scales that provide a continuum from negative to positive values indicate that the concept measured by the scale is bipolar in nature, whereas scales that present only positive values indicate a unipolar conceptualization.
Item Response Theory

When item format is polytomous, partial credit models are useful in providing estimates of the psychological distance between each set of ordinal categories (Fox & Jones, 1998). The ordered responses are called category scores, and each successive score point indicates an additional correct operation or higher level of the attribute being measured. The Rasch partial credit model evaluates each pair of adjacent responses as if they were dichotomous, identifying a transition point, or transition location parameter, between each pair that divides individuals on the basis of the levels of attribute being measured. The transition location is the point at which the probability of selecting either of the two options is equal. The probability of attaining each category score can be graphed as an option response function (ORF). The generalized partial credit model is an extension of the partial credit model that permits a discrimination parameter to be estimated separately for each item (de Ayala, 2009).

In the one-parameter rating scale model, the transition points represent thresholds at which the probability of selecting the next higher response is attained. The rating scale model requires the same number of response categories for all items in the item set and the thresholds are estimated simultaneously for all items in the set. The two-parameter graded response model is an extension of the rating scale model in which the response categories are computed as a series of cumulative dichotomous steps contrasting the score below a category with scores at or above that category (de Ayala, 2009). With this model, items may have different numbers of response options and different discrimination parameters. A benefit of the graded response model over the rating scale or partial credit models is that the ordering of response options is preserved. However, any of the partial credit or graded response models may be used with Likert responses.
Method

Instrument

The Intercultural Student Attitude Scale (ISAS) is a 35-item survey that measures the cross-cultural attitudes of undergraduate students studying abroad. The ISAS has been validated with published cross-cultural scales and personality measures (Shaftel & Shaftel, 2010). The ISAS was presented via an online survey engine to four different groups of university students with four different sets of response options. To maximize sample size, two groups of students were assessed during November 2011 and two additional groups in February 2012. A five-point standard Likert rating scale was used with the first group. Response options ranged from strongly disagree to strongly agree. The second was a seven-point scale (not at all like me, not like me, not much like me, neutral, somewhat like me, like me, just like me). The third group used a nine-point numerical scale with verbal anchors only at the endpoints (completely untrue, completely true). Finally, a 0-100 visual analog scale marked in tens (i.e., 0, 10, 20, etc.) was presented to the fourth group. Respondents clicked on the point on the line that represented their level of agreement. Each of these scales was selected because it is commonly used and familiar to respondents, with the possible exception of the 0-100 visual analog scale. One earlier dataset was included in the analysis to contrast the six-point verbally anchored scale that has been used with the ISAS throughout its development. The six-point scale options include definitely not true, not true, tends to be not true, tends to be true, true, definitely true.

Participants

Participants were 1554 college students who completed the questionnaire for extra credit in a large undergraduate business course at a Midwestern university. Demographic information obtained from the respondents is included in Table 1.
Analyses

Scores for items that measured negative interpretations of the attribute, even if worded positively, were reversed so that higher scores indicated higher levels of the attribute. Records for which there was no variability in response (e.g., the student answered every question with the highest value) were removed. Total scores were computed as the mean of all items in the scale so that the distributions could be compared among the different scales. Histograms for individual items and total scores were prepared for comparison with the IRT output. Reliability was assessed using coefficient alpha, and point-biserial correlations for each item with the total score were obtained.

Unidimensional two parameter logistic graded response models were computed using IRT Pro to evaluate difficulty and discrimination parameters for each item. The beta values in the output represent the theta values, or levels of the measured attribute, at which the probabilities of choosing two adjacent response categories are equal. Theta values are distributed as z-scores with a mean of 0 and standard deviation of 1. The respondent with \( \theta \)-level = \( \beta_i \) has an equal probability of choosing response categories \( i-1 \) and \( i \). Distances between each of the betas for every item on each scale were calculated, and the standard deviation of those distances was used as a measure of the variability of the intervals between each step of the ordinal scale. Visual inspection of the ORF curves revealed the number of categories that provided unique information for each item. IRT output provided the transition points, or betas, between response categories for each item. The differences between each pair of betas were computed and the mean difference was found for each item on each scale. The average of the mean differences for each item gives a measure of the relative size of the distances between transition points for each scale. The variability of the distances between transition points was computed as the standard
deviation of the transition point differences for each item. The mean standard deviation of the differences between betas provides a measure of how closely the data for each scale approached interval-level measurement.

Poorly performing items were identified as those with only one response option providing unique item information between -3 and 3 or with zero or one transition location or beta value between -3 and 3. These selection criteria were chosen because an item that only offers one usable response for all individuals with ability levels from -3 to 3, regardless of the number of beta values, doesn’t add anything to the scale. Similarly, zero or one beta value between -3 and 3 shows that most of the response options are too extreme for usefulness, even if they provide unique information. While beta values precisely define the probability of the corresponding response options, many response options may not add unique information to the measurement of the item, and thus the item may perform poorly even if several beta values are within the critical measurement range of -3 to 3 standard deviations around the mean of the measured ability.

The frequencies for the 101-response scale did not generally follow the normal curve. To improve the normality of the data, that dataset was recoded into a new dataset with 10 response categories (0-9=1, 10-19=2 … 90-100=10). Other problems analyzing the 101 point scale prevented it from being included in these preliminary analyses.

**Results**

Table 1 displays summary data about the various rating scale options. The optimal value for each category is shown in bold italics, though it is apparent that very similar values are also available in adjacent rating scale lengths. Internal consistency, as measured by coefficient alpha, suggests that rating scale length has little effect on the homogeneity of scale items in this study, with all values at or approaching .90. Similarly, the mean point-biserial correlation is fairly
consistent for all the scales, with the 6, 7, and 9 point options displaying slightly higher overall values. It is important to note that both coefficient alpha and point-biserial correlation are techniques designed to be used with interval, not ordinal, data (Howe, 2006), and thus may be misrepresent these data to some extent.

The number of poorly performing items suggests that increasing the number of response options tends to improve the performance of individual items, expanding some weak items into adequate measures of the construct for this survey instrument. The 5 and 6 point scales, in particular, have several more bad items than the 7 point and longer response scales. Six items on the 5 point scale have only one beta value between -3 and 3. Fewer items have this problem on the longer response lengths. The same poor item has no beta value in the critical range on the 7 and 9 point scales, and this item performs poorly in other ways on every scale, demonstrating that this is an item problem rather than a scale problem.

The mean distance between beta values for all 35 items becomes smaller as the number of response options increases, as might be expected. However, diminishing returns are evident as the length of the rating scale increases. Using 5 response options offers a mean distance between transition locations of over 3 points, or standard deviations, on the theta scale, which is an extremely coarse differentiation of individual differences for the attribute measured by the ISAS. Increasing the number of response options to 10 only reduces that value to 1 1/3 theta standard deviations, which still cannot be said to provide a fine-grained distinction.

The standard deviation of the distances between beta values offers a sense of the variability for the different rating scale lengths. The 5 point scale shows a standard deviation of beta differences of slightly more than one theta point. The variability of distances between transition locations becomes smaller as the number of response options increases, with the 10
point scale having an average standard deviation of less than half a theta point. However, all of
the scales with more than 5 response categories have much lower variability than the 5 point
scale. The mean, minimum, and maximum number of beta values within the -3 to 3 range also
contributes an idea of the variability among items. The mean number of transition locations
obviously increases as rating scale options increase.

For each rating scale option, there was at least one item with either zero or one beta value
in the critical measurement range of -3 to 3 standard deviations around the theta mean. The
number of items with the lowest or highest number of beta values is shown in parentheses. In this
study, one particular item was frequently the source of this poor measurement. In contrast, the
maximum number of betas within the key measurement range increased as the number of
response options increased, though few items exhibited the maximum number of available beta
values within that range. Even with a larger number of beta values between -3 and 3, however,
not all response categories may contribute unique information to item measurement.

In order to evaluate the number of response options that provided unique information for
the measurement of each item, visual inspection of the ORFs was conducted. The mean number
of response categories with unique measurement information shows an interesting pattern of
increase from just over 3 to about 4½ at the 7 and 9 point scales, diminishing to under 4 at the 10
point scale. This particular result indicates that greater numbers of response options does not lead
to greater use of those options overall. The minimum and maximum number of response
categories with unique information demonstrates that one or more items on each scale use only 1
or 2 options while a small number of items contribute a larger number of meaning response
options. Again there is a pattern of diminishing returns, with the maximum number of effective
response categories dropping off as the length increases to more than 9 rating scale points.
Histograms for the total score distributions, which were computed as the mean of all 35 items on the ISAS, for each rating scale are shown in Figures 1-5. Table 2 summarizes the distributional properties of each scale. All of the total score distributions are essentially normal, with only very mild skew or kurtosis in two instances. Visual inspection of the distributions reveals that the 5 and 6 point scales appear to have better symmetry and modality about the mean, though the 9 point scale is also not bad.

**Discussion**

The IRT analysis performed for this project revealed some previously unknown characteristics of the ISAS rating scale. Some items performed poorly with different numbers of scale points, even though all items had previously been deemed adequate with classical test theory analyses. Most interesting are the items that were ineffective with some rating scales but effective with others. These items are shown in Table 3 and ORFs are included in the appendix. An example is item 1, which had extreme transition locations on the 5 and 6 point scales but better values as the number of scale points increased to 9. Two items in particular, 6 and 20, accounted for many of the problems. According to these IRT results, these items were not particularly effective even when they did not meet the subjective criteria used in this project for bad items. However, the current results do not necessarily imply that these items should be removed from the scale if they are useful for other purposes, such as measuring the attitudes of students who study abroad as opposed to the convenience samples assessed for this study.

While several descriptive statistics were used to summarize the different rating scale lengths, these results did not all lead to the same conclusion about the optimal number of response categories. Some scale attributes, such as the mean distance between transition locations and the mean number of response categories that provided unique measurement
information, showed diminishing returns or even reversed as scale length increased. These features have not previously been evaluated with different numbers of response options. In contrast, internal consistency as measured with coefficient alpha was virtually unchanged across the different scale responses, which is consistent with most previous research (Aiken, 1983; Bendig, 1954; Matell & Jacoby, 1971).

Conclusions

This study provides only preliminary results of these rating scale comparisons. First, due to software limitations, the ORFs were computed for each scale in two parts. When all 35 items for each scale can be included in a single analysis, the ORFs and other item features may change slightly. For the same reason, the 101 point scale could not be adequately analyzed and results could not be prepared for inclusion in this paper.

Furthermore, these results are not generalizable because the characteristics of each item are unique to the ISAS. Poor items on the ISAS contributed to problems at each rating scale length. Therefore, additional graded response IRT analyses of other surveys would add to the body of knowledge that would inform choices about the number of response options for different purposes. Numbers and types of response categories may also be dictated by item content, so a simple decision rule about the “correct” number of options would probably not be feasible in most instances. Nonetheless, confirmation or disconfirmation of these results in terms of the size and variability of intervals between response option choices may lead to some convergence on recommendations for scale lengths in various situations.

Limitations of this study include the sample size for each rating scale. While sufficient for IRT analyses, sample size was limited by the requirement that separate samples complete the
survey with different numbers of scale points; even so, these samples were obtained over several semesters. Further research with other surveys and respondents may help address this problem.

This study adds to the literature on development of rating scales through IRT analyses of polytomous ordinal data using the graded response model. Intervals between response categories can be described precisely, evenness and spacing of response options can be graphically represented, and direct comparisons between rating scales can be made. With additional research, researchers may be able to make better decisions about the numbers and types of scale points for specific rating scale purposes.
References


doi:10.1177/001316448304300209


doi:10.1207/s15324818ame0902_4


doi:10.3102/00346543071001105

Howe, K. J. (2006, August). The measurement of composite variables from multiple indicators: Applications in Quality Assurance and Accreditation Systems – Childcare. Background


Table 1

**Characteristics of Scales with Different Numbers of Rating Scale Response Options**

<table>
<thead>
<tr>
<th>Characteristics of Scale</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>9</th>
<th>10</th>
</tr>
</thead>
<tbody>
<tr>
<td>N of subjects</td>
<td>224</td>
<td>634</td>
<td>239</td>
<td>212</td>
<td>245</td>
</tr>
<tr>
<td>Percent male</td>
<td>67%</td>
<td>53%</td>
<td>65%</td>
<td>66%</td>
<td>58%</td>
</tr>
<tr>
<td>Percent European American</td>
<td>77%</td>
<td>75%</td>
<td>76%</td>
<td>75%</td>
<td>66%</td>
</tr>
<tr>
<td>Coefficient alpha</td>
<td>.89</td>
<td><strong>.90</strong></td>
<td><strong>.90</strong></td>
<td><strong>.90</strong></td>
<td>.87</td>
</tr>
<tr>
<td>Number of poor items*</td>
<td>6</td>
<td>5</td>
<td>2</td>
<td>3</td>
<td><strong>1</strong></td>
</tr>
<tr>
<td>Mean point-biserial correlation</td>
<td>.42</td>
<td><strong>.45</strong></td>
<td><strong>.45</strong></td>
<td>.44</td>
<td>.39</td>
</tr>
<tr>
<td>Mean distance between betas</td>
<td>3.30</td>
<td>1.98</td>
<td>2.06</td>
<td>1.57</td>
<td><strong>1.33</strong></td>
</tr>
<tr>
<td>Mean SD of distance between betas</td>
<td>1.05</td>
<td>.45</td>
<td>.55</td>
<td>.59</td>
<td><strong>.42</strong></td>
</tr>
<tr>
<td>Mean # of betas between -3 and 3</td>
<td>2.37</td>
<td>3.03</td>
<td>3.94</td>
<td>5.06</td>
<td><strong>5.97</strong></td>
</tr>
<tr>
<td>Lowest # of betas between -3 and 3^</td>
<td>0 (1)</td>
<td>1 (3)</td>
<td>0 (1)</td>
<td>1 (2)</td>
<td>0 (1)</td>
</tr>
<tr>
<td>Highest # of betas between -3 and 3^</td>
<td>4 (5)</td>
<td>5 (3)</td>
<td>6 (7)</td>
<td>8 (3)</td>
<td><strong>9 (6)</strong></td>
</tr>
<tr>
<td>Mean # of response options with unique info</td>
<td>3.17</td>
<td>3.54</td>
<td>4.43</td>
<td><strong>4.45</strong></td>
<td>3.71</td>
</tr>
<tr>
<td>Lowest # of response options with unique info</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>Highest # of response options with unique info</td>
<td>5</td>
<td>5</td>
<td>7</td>
<td><strong>8</strong></td>
<td>7</td>
</tr>
</tbody>
</table>

* A poor item is one with only one effective response category or zero or one transition location between -3 and 3.
^ The number of items with this characteristic is shown in parentheses.

Note: The best values for each category are shown in bold italics.
Table 2

*Characteristics of Total Score Distributions for Each Scale Length*

<table>
<thead>
<tr>
<th>Total Score Distribution</th>
<th>5 points</th>
<th>6 points</th>
<th>7 points</th>
<th>9 points</th>
<th>10 points</th>
</tr>
</thead>
<tbody>
<tr>
<td>N</td>
<td>224</td>
<td>642</td>
<td>238</td>
<td>212</td>
<td>245</td>
</tr>
<tr>
<td>Mean</td>
<td>3.71</td>
<td>4.43</td>
<td>5.00</td>
<td>6.30</td>
<td>7.16</td>
</tr>
<tr>
<td>Skew</td>
<td>Normal</td>
<td>Normal</td>
<td>Normal</td>
<td>Normal</td>
<td>Mildly negative</td>
</tr>
</tbody>
</table>
| Kurtosis                 | Normal   | Normal   | Mildly negative | Normal   | Normal
Table 3

Comparison of Items Whose Effectiveness Varied by Rating Scale Length in Terms of Number of Response Options Contributing Unique Information (#RO) and Number of Transition Locations Between -3 and 3 (#TL)

<table>
<thead>
<tr>
<th>Item</th>
<th>5 points</th>
<th>6 points</th>
<th>7 points</th>
<th>9 points</th>
<th>10 points</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>#RO</td>
<td>#TL</td>
<td>#RO</td>
<td>#TL</td>
<td>#RO</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>5</td>
</tr>
<tr>
<td>6</td>
<td>3</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>16</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>17</td>
<td>1</td>
<td>0</td>
<td>2</td>
<td>1</td>
<td>5</td>
</tr>
<tr>
<td>19</td>
<td>1</td>
<td>0</td>
<td>2</td>
<td>1</td>
<td>5</td>
</tr>
<tr>
<td>20</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>27</td>
<td>3</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>3</td>
</tr>
<tr>
<td>31</td>
<td>2</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>3</td>
</tr>
</tbody>
</table>

Note: Values meeting criteria for poor fit are noted in bold italics.
Figure 1. Total score distribution for 5 point scale.
Figure 2. Total score distribution for 6 point scale.
Figure 3. Total score distribution for 7 point scale.
Figure 4. Total score distribution for 9 point scale.
Figure 5. Total score distribution for 10 point scale.