Middle category or survey pitfall - Using Rasch modeling to illustrate the middle category measurement flaw

Kelly D. Bradley¹
University of Kentucky

Jessica D. Cunningham
Western Carolina University

Kathryn S. Akers
University of Kentucky

Nikki Knutson
University of Kentucky

¹ Use Dr. Bradley as author of contact, via kdbrad2@uky.edu
Abstract

The quality of the instrument used in the measurement process of survey data is fundamental to successful outcomes. Content and structure considerations are at the lead of the discussion, but the rating scale is just as important. Specifically for likert-type questionnaires, the words used to describe rating categories and the placement of the neutral or not sure category is at the core of this measurement issue. This study utilizes the Rasch model to assess the quality of the measurement instrument and structure of the measurement scale for a typical data set collected at a higher education institution. Highlights include an illustration of the importance of category placement and an evaluation of the use of a middle category for likert-type survey data.
Middle category or survey pitfall - Using Rasch modeling to illustrate the middle category measurement flaw

The use of surveys in educational research remains one of the most popular research methodologies for graduate studies and published papers. With so many research studies utilizing survey research methods, it is increasingly important that survey instruments are functioning the way they are intended and measuring what they claim. Bond and Fox (2004) write, “Operationalizing and then measuring variables are two of the necessary first steps in the empirical research process. ... Thus, interpretation of analyses can only be as good as the quality of the measures.” The quality of the instrument used in the measurement process must play a fundamental role in the analysis of the data collected from it. It is important to begin at the level of measurement and to identify weaknesses that may limit the reliability and validity of the measures made with the survey instrument.

The objectives of this study are to assess the quality of the measurement instrument and the structure of the rating scale. Specifically, this study examines the inclusion, or exclusion, of a middle category for likert-type survey responses. This middle category can have many different titles such as: neutral, not sure, or neither category. The meaning of this category is often unclear, and researchers have hypothesized that respondents may be interpreting this category in a variety of different ways. Hence, it is reasonable to believe these responses are different from the intent of the instrument.

Rationale and Background

Many researchers include a middle category as a way to make survey respondents feel comfortable; however, by allowing the respondent to be noncommittal and choose a response
that is physically in the middle of a scale, essentially, the specificity of measurement is being diminished, or even washed away. Potentially, this creation of safety to respond is leading to a loss of information.

*Middle Category*

When employing a “neutral” or “unsure” response option on a survey, researchers should be mindful that this scale construction could have major implications on their survey data (Ghorpade & Lackritz, 1986). Selecting a neutral response option can be deciphered to suggest varying intents by the survey respondent. Choosing the “neutral” response may be interpreted as the midpoint between two scale anchors, such as a positive and a negative (Ghorpade & Lackritz, 1986). Selecting the “neutral” response option could also represent that the respondent was not familiar with the question or topic at hand and as a result, was not sure how to answer this particular item. A further interpretation is that this response option indicates that the respondent does not have an opinion to report or they are simply not interested in the topic. Research has shown that the presence of a “no opinion” or “don’t know” option reduces the overall number of respondents that offer opinions (Schaeffer & Presser, 2003). One explanation for this finding is that it is easier for a respondent to simply select the “neutral” or “no opinion” response category as it easier than contemplating over the other response options (DeMars & Erwin, 2004). Respondents may select the “neutral” category if the survey items are difficult or if the respondent has low motivation to complete the survey (DeMars & Erwin, 2004).

While respondents may view response categories of “neutral”, “not sure” and “no opinion” to be one in the same, these options are in fact dissimilar and these response options should be used with intention. Researchers apply the “neutral” category to indicate that a
respondent is declaring the middle position between two points while “no opinion” or “don’t know” are intended to reflect a lack of opinion or attitude towards a statement (DeMars & Erwin, 2004). A further response option used as a middle category in survey research is “unsure” or “undecided”, these options are intended to be used as an option when a respondent is having difficulty selecting between the other available responses (DeMars & Erwin, 2004).

Even though these neutral response categories are frequently used on surveys administered in postsecondary educational settings, the literature confirms that how these responses should be scored is largely undetermined (DeMars & Erwin, 2004). One justification for this finding is that neutral response options do not affect all surveys equally and therefore one method for working with neutral response options is not generalizable as researchers need to consider the survey construct (DeMars & Erwin, 2004). If researchers use a middle category, they must decide how these responses will be scored. In survey research, using neutral response options are often manipulated in an overly simplistic manner (Ghorpade & Lackritz, 1986).

Rasch versus Classical Test Theory Approach

As noted in Smith (2000), the classical test theory model, sometimes called the true score model, has deficiencies. This approach requires complete records to make comparisons of items on the survey. Even if this is attained, the issue of sample-dependence between estimates of an item’s difficulty to endorse and a respondent’s willingness to endorse surface. Moreover, the estimates of item difficulty cannot be directly compared unless the estimates come from the same sample or assumptions are made about the comparability of the samples. Another concern with the classical test theory approach is that a single standard error of measurement is produced for the composite of the ratings or scores, inadequate at best.
The Rasch model, introduced by Georg Rasch (1960), addresses many of the weaknesses of the classical test theory approach. It yields a more comprehensive and informative picture of the construct under measurement as well as the respondents on that measure. Specific to rating scale data, the Rasch model allows for the connection of observations of respondents and items in a way that indicates a person endorsing a more extreme statement should also endorse all less extreme statements, and an easy-to-endorse item is always expected to be rated higher by any respondent (Wright and Masters, 1982). Information about the structure of the rating scale and the degree to which each item contributes to the construct is also produced. The model provides a mathematically sound alternative to traditional approaches of survey data analysis (Smith, E., 2000; Wright, 1997).

Physical placement of the middle item can have implications for the respondent and analysis of survey responses. A scale that reads strongly disagree to disagree to agree to strongly agree is assumed to increase with each step of the scale, agreeing with the item more and more. When a response such as neutral or not sure is inserted physically in between disagree and agree, it can no longer be assumed that the categories are arranged in a predetermined ascending or descending order. Another issue that researchers must tackle when dealing with a middle category in a likert-type survey scale is the interpretation of the middle category. What does it mean for a respondent to select not sure, neutral or neither on an item that the respondent should and probably does have an opinion about? The answer to this can have profound impacts on the analysis of the responses.

Here, the issue of whether or not to include a middle category, as well as placement of the middle category, for likert-type survey instruments is examined. This study has practical and
methodological implications. It serves as an assessment of the survey itself by ensuring the survey is functioning as it was intended. It can also benefit survey developers by giving insight into possible revisions for future rounds of data collection. Methodologically, this study serves as a framework for educational researchers developing survey instruments and analyzing rating scale data.

Methods

This study utilizes the Rasch model to assess the measurement instrument and specifically, the structure of the measurement scale of a typical data set collected in higher education settings.

Instrumentation

The entire instrument was comprised of 29 items including six items for questions pertaining to student demographic information. The scale examined in this study was a five-point rating scale; 1=Strongly Disagree, 2=Disagree, 3=Not Sure, 4=Agree, 5=Strongly Agree. The survey item asked respondents to indicate their level of agreement or disagreement with a series of 13 statements related to academic readiness. The instructions clearly outline that there are no correct or incorrect responses and asks students to select the response that best reflects their values and beliefs. The statements were adapted from existing scales on the topics of student self-efficacy and procrastination, both of which are shown to be indicators of student college readiness in existing literature.

Data Set

The response frame included responses from 1,665 first year students at a large research university in the southeast United States. The Office of Institutional Effectiveness sent out a broadcast email with an imbedded web-survey link to all first year students (as determined by the
University Registrar) at the university in the fall 2009 semester. Instructions read that the survey would take approximately 15-20 minutes to complete. Students were informed that although student identification numbers were collected, all responses would be aggregated with the goal of confidentially. Student identifiers ensure that all students who responded were in fact, first year students at the university. Once this was verified, student identifiers were removed from the data set for all future analyses. The data was housed in an excel file and stored on a secure web server. This study was conducted with the institution’s IRB approval.

Analysis

A Rasch model was employed, because it uses the sum of the item ratings simply as a starting point for estimating probabilities of those responding. Item difficulty is considered the main characteristic influencing responses, and it is based on the ability to endorse a set of items and the difficulty of a set of items (Linacre, 1999). In general, people are more likely to endorse easy-to-endorse items than those that are difficult to endorse. People with higher willingness-to-endorse scores are more agreeable than those with low scores. Thus, it is a typical rating scale.

A rating scale model was applied to test the overall data fit to the model by using the software package, WINSTEPS version 3.57.4 (Linacre, 2005). Missing data were treated as missing, as the Rasch model has the ability to deal with such entries without imputing estimates or deleting any portion of the data. Of the 1,665 respondents who submitted responses to a web-based survey, 1,595 respondents were measured on the 13 rating scale items this study examined. 140 responses were coded as missing. Once fit in this fashion, the next step was to delete the response category Not Sure and the model was re-run. Finally, the data were put through a final analysis with the Not Sure category added to the end of the scale. These analyses provided for
three scenarios – the original 5-point scale, a 4-point scale with the middle category removed, and a 5-point scale where the not sure has been lifted and reinserted at the end.

The model chosen for this analysis is the Polytomous Rasch model or rating scale model (Andrich 1978), an extension of the general Rasch model. It is specified as:

$$\log \left( \frac{P_{nj}}{P_{nj-1}} \right) - B_n - D_i - F_j$$

where $P_{nj}$ is the probability that person $n$ encountering item $i$ is observed in category $j$, $B_n$ is the “ability” or rater-severity measure of person $n$, $D_i$ is the difficulty-to-endorse measure of item $i$, and $F_j$ is the “calibration” measure of category $j$ relative to category $(j-1)$ (Linacre, 2004).

Results and Discussion

To investigate measurement flaws related to the middle category, a logical place to begin the discussion is to inspect the fit and function of the rating scale itself. Table 1 includes the observed average measures of the rating scale categories for all items.

Table 1: Observed Average Measures across Analyses

<table>
<thead>
<tr>
<th></th>
<th>Middle Category in Middle</th>
<th>Middle Category to End</th>
<th>Middle Category Missing</th>
</tr>
</thead>
<tbody>
<tr>
<td>Strongly Disagree</td>
<td>-.53</td>
<td>-.04</td>
<td>-.55</td>
</tr>
<tr>
<td>Disagree</td>
<td>-.21</td>
<td>.27</td>
<td>-.24</td>
</tr>
<tr>
<td>Not Sure</td>
<td>.32</td>
<td>.82</td>
<td></td>
</tr>
<tr>
<td>Agree</td>
<td>1.04</td>
<td>.58</td>
<td>1.04</td>
</tr>
<tr>
<td>Strongly Agree</td>
<td>2.11</td>
<td>.82</td>
<td>2.25</td>
</tr>
<tr>
<td>Missing</td>
<td>.43</td>
<td>.25</td>
<td>.45</td>
</tr>
</tbody>
</table>
According to Linacre (2004), the average observed measures should increase monotonically as the rating scale increases. The rating scales with the middle category included in the middle and coded as missing follow this sequence. However, the recoded middle category to the end does not function as well with the observed average measures not increasing from “strongly agree” (4) to “not sure” (5). In essence, if the middle category is included at the end of a survey rating scale rather than the middle it should remain coded as the middle rating scale category. Since the observed average measures increase for both the middle category in the center and the middle category removed, the step calibrations are span over a greater distance of logits with the middle category removed and are thus preferred from a measurement perspective.

Figure 1 is an example of a Rasch estimated category probability curve for the original 5-point scale. The x-axis represents what is being measured, here freshman preparedness, as defined by the question on the survey being analyzed. The y-axis represents the probability of responding to any category, ranging from 0 to 1. For example, looking at category 1, strongly disagree, the likelihood of a person responding strongly disagree decreases as their level of preparedness increases. Theoretically, each response category will peak at some point on the graph; however, in the probability curve in Figure 1 it is clear that category 3, not sure does not have a true peak. Figure 2 shows the same probability curve, with category 3 removed.
### Figure 1 Category probability curves for the 5 response option categories

In general, a quick comparison between Figures 1 and 2 shows that all categories peak higher when 3 is removed. The probability curve where the middle category is removed is optimal in comparison to the others, since the probability curve of each rating category peaks within the range of possible respondent estimates (see Figure 2). The rating category *not sure* does not peak in the probability curves for Figure 1 with the middle category in the center of the rating scale.
Figure 2 Category probability curves for the 4 remaining response option categories

The conclusion here may be that when category 3 is included, it is washing out the measurements for the other categories. For example category 2, strongly disagree peaks strong and more prominently when category 3 is removed. The rating category strongly agree is subsumed in Figure 3 by the adjacent probability curves along the rating scale. Thus, the middle category removed is optimal from a measurement standpoint from this part of the analysis.
One way to examine the use of item categories using the Rasch rating scale model is by examining category probability curves. Average measures increasing monotonically with the rating scale for each item is also an indication that the rating scale is functioning as expected for each item in the survey instrument. For the analysis with the middle category centered in the rating scale, items 9 and 13 had a disordered average measure sequence within the rating scale categories. For item 9, the rating category disagree was expected to observe an increased average respondent measure and empirically showed the opposite, a decrease. Similarly, item 13 showed a decrease in average respondent measure from disagree to not sure. Questions 9 and 13 remained problematic with non-increasing average measures for strongly disagree when the
middle category was removed from the analysis. In the analysis with the middle category recoded to the end of the rating scale, a monotonic increase was seen in all items moving along the rating scale. The conclusion here is a classic example of how the middle category may be interpreted differently depending upon the survey item.

The item hierarchy, which can also be viewed as the endorsability of the survey items, is also an important aspect to consider in the inspection of whether a middle category is impacting the measure. Table 2 indicates the item estimates for each item based upon the three types of analysis, considering inclusion, exclusion, or ordering of the middle category.

Table 2. Item Estimates

<table>
<thead>
<tr>
<th>Item Number</th>
<th>Middle Category in Center</th>
<th>Middle Category Removed</th>
<th>Middle Category to End</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>.46</td>
<td>.45</td>
<td>.08</td>
</tr>
<tr>
<td>2</td>
<td>.47</td>
<td>.52</td>
<td>.28</td>
</tr>
<tr>
<td>3</td>
<td>-.22</td>
<td>-.17</td>
<td>-.05</td>
</tr>
<tr>
<td>4</td>
<td>.38</td>
<td>.35</td>
<td>-.11</td>
</tr>
<tr>
<td>5</td>
<td>-.42</td>
<td>-.37</td>
<td>-.01</td>
</tr>
<tr>
<td>6</td>
<td>.79</td>
<td>.82</td>
<td>.35</td>
</tr>
<tr>
<td>7</td>
<td>.34</td>
<td>.28</td>
<td>-.09</td>
</tr>
<tr>
<td>8</td>
<td>.59</td>
<td>.62</td>
<td>.05</td>
</tr>
<tr>
<td>9</td>
<td>-.53</td>
<td>-.52</td>
<td>-.06</td>
</tr>
<tr>
<td>10</td>
<td>-.11</td>
<td>-.08</td>
<td>-.09</td>
</tr>
<tr>
<td>11</td>
<td>.48</td>
<td>.51</td>
<td>.11</td>
</tr>
<tr>
<td>12</td>
<td>-.51</td>
<td>-.57</td>
<td>-.21</td>
</tr>
<tr>
<td>13</td>
<td>-1.71</td>
<td>-1.82</td>
<td>-.25</td>
</tr>
</tbody>
</table>
The average item measure is set at zero. The standard deviations were 0.66, 0.68, and 0.17 respectively. All items remained on the same “side” of the continuum from the mean item measure across all analyses with the exception of items 4 and 7. Item 4 and 7 were above the item mean measure except when the middle category was recoded to the end of the rating scale. The hierarchy of the items can be inspected graphically using an item map displayed in Figures 4, 5 and 6. Here, probabilities based on each respondents choice for each item places each item along this continuum where the most likely to be endorsed items are at the bottom while the most difficult to endorse items fall towards the top. A disordering of the item hierarchy would indicate the construct being measured is affected by the inclusion or displacement of the middle category.
Figure 4 Item endorsability for middle category in center
Figure 5 Item endorsability for middle category removed
Figure 6 Item endorsability for middle category moved to end of rating scale

One important distinction would be to notice the low variability in item measures when the middle category is recoded to the end of the rating scale (see Figure 6). Secondly, the disordering of the items differs in Figure 6. As well, in comparison to the hierarchies are approximately the same between Figure 4 and 5. It appears that the middle category coded at the end of the rating scale presents a different construct being measured by the survey instrument.
Conclusion

Within the field of education, the development of instruments to assess affective domain constructs has been a problematic area (Aiken, 1996; Martin, 1983). Surveys are the most common example of self-reported data collection and continue to be one of the most popular research methodologies for graduate studies and published papers in education (Aiken, 1988; Babbie, 1992; Gay, 1981). Even so, the efficiency and effectiveness of the instrument as a measurement tool is often overlooked or underemphasized.

As noted by Sampson and Bradley (2003) the classical test theory model produces a descriptive summary based on statistical analysis, but it is limited if not absent of the capability to assess the quality of the instrument. It is important to begin at the level of measurement and to identify weaknesses that may limit the reliability and validity of the measures made with the instrument. As indicated in the study, Rasch analysis tackles many of the deficiencies of the classical test theory model in that it has the capacity to incorporate missing data, produces validity and reliability measures for person measures and item calibrations, measures persons and items on the same metric, and is person and sample-free.

A methodological framework for educational researchers developing survey instruments and analyzing rating scale data is offered. In analyzing results collected via a survey instrument, it is presumed the respondents have an accurate perception of the construct, rate items according to reproducible criteria, and accurately record their ratings within uniformly spaced levels. In fact, as noted in Wright (1997), ratings are simply responses based on fluctuating personal criteria, the responses are not always interpreted as intended or recorded correctly, and these ratings are ordinal so they do not add up to measures. Rasch analysis produces measures,
provides a basis for insight into the quality of the measurement tool and provides information to allow for systematic diagnosis of misfit. This study illustrates and highlights many issues with the inclusion and placement of a middle category through a measurement lens. The take away point is when utilizing ambivalent response options on a survey, researchers should examine response patterns to these items within their survey constructs and datasets before interpreting and reporting their result.
References


