Making Meaningful Measurement in Survey Research: The Use of Person and Item Maps

Kenneth D. Royal, Ph.D.

American Board of Family Medicine
Psychometrics Department

September, 2009

Making Meaningful Measurement in Survey Research: The Use of Person and Item Maps

Perhaps the greatest limitation of higher education research today pertains to quality measurement. In 1959, S.S. Stevens provided the widely cited definition of measurement in the social sciences. That is, *measurement is the assignment of numerals to events or objects according to rule* (p. 25). Unfortunately, many who read this work ignored his latter statement in the same text:

> When operations are available to determine only rank order, it is of questionable propriety to compute means and standard deviations... If we want to interpret the result of averaging a set of data as an arithmetic mean in the usual sense, we need to begin with more than an ordinal assignment of numerals. (p. 29)

Distinguishing the difference between ordinal and interval scales is essential to quality measurement in the social sciences, especially survey research. Most survey researchers typically incorporate some form of an ordinal scale to coincide with the measurement of survey items. Then, the data are treated as if they were interval and linear statistical techniques are applied. Unfortunately, most survey researchers fail to realize most rating scales simply distinguish rank among response options. That is, one response option indicates more or less of something than the other options. These scales are not interval measures and should not be treated as such.

Consider the following example of survey items provided by Bond and Fox (2001). A sample of grade school children was asked two questions:

1) I am afraid that I will make mistakes when I use my computer.
2) I am so afraid of computers I avoid using them.

A rating scale containing the following response options are provided: *Strongly Disagree, Disagree, Neutral, Agree,* and *Strongly Agree.* In theory, we expect our rating scale to look something like this:

	Less Anxious				More Anxious
1) Mistake	SD	D	N	A	SA
2) Avoid using	SD	D	N	A	SA

But, in reality, it may look something more like this:

	Less Anxious				More Anxious
1) Mistake		SD	D N	A	SA
2) Avoid using	SD D N A SA				

Under the classical approach our problems only multiply when we add values (or scores) to these data. Hypothetical results might provide a mean of 4.0 with a standard deviation of .8 for question #1, and a mean of 2.0 with a standard deviation of .8 for question #2. What can one *truly* say about these items given simply a mean and standard deviation for each? Can one really make any meaningful inferences about individual respondents, or their responses relative to other respondents? Typically when interpreting the results, one would compare the two items and say that people generally agreed with the former statement more so than the latter. One would then look at the mean scores and standard deviations and try to get a sense of the average level of agreement/disagreement indicated by these scores. Further clouding this picture, one would try to imagine how standard deviations affect all this. (This, of course, is excluding any discussion about sampling strategy, whether the sample data are representative of the population, and other methodological concerns!) What is learned is that adding another item suddenly imposes additional problems for ones interpretation of results.

In the scenario given above yet another erroneous assumption is made. This time, it is a failure to realize that all survey items are not equally important. Given the example provided, the second item clearly demonstrates a greater fear of computers. Suppose students generally disagreed (rating = 2.0) with the statement “I am so afraid of computers that I avoid using them”, yet also generally disagreed (rating = 2.0) with the statement “I am afraid that I will make mistakes when I use my computer”. Does it really make any intuitive sense to score both of these items as 2.0 and treat them as though they were of equal importance?

In the fields of assessment and institutional research these mistakes are made all the time. Most are unaware that these techniques fall under a methodological approach called "classical test theory" (CTT). Although CTT techniques have their purposes, they are largely ineffectual in the arena of survey research. In particular, CTT techniques have long been criticized for several additional and important reasons. For example, error across measurement units is independent and uncorrelated (Becker, 2001). Missing data are problematic and CTT has no way of appropriately addressing the issue (Moulton, 2009). Further, CTT approaches are sample dependent, require larger samples and necessitate representative data (Bond & Fox, 2001; Bunderson, 2000; Hambleton, Swaminathan, & Rogers, 1991). Fortunately, there are solutions to these problems.

Some sixty years after Spearman introduced CTT (Spearman, 1904), scholars began to re-examine CTT and its assumptions and began to develop new models with stronger theoretical underpinnings. The more theoretically sound solution came to be known as Item Response Theory (IRT). Whereas CTT focuses on test level information, IRT focuses on item-level information. In other words, IRT focuses on the interactions between individual persons and items, as the model suggests, each affects the other. Although there are numerous IRT models, the one-parameter logistic model, known as the Rasch model, is generally regarded as the gold standard. In addition to the Rasch model being able to convert ordinal data to interval measures, the model hosts a number of advantages over CTT approaches. Bradley and Sampson (2005) have eloquently summarized many of those advantages stating:

> Whereas the classical model produces a descriptive summary based on statistical analysis, it is limited, if not absent, in the measurement capacity. In contrast, Rasch measurement tackles many of the deficiencies of the classical test 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 not dependent on the particulars of the sample. Applications of the Rasch model allow the researcher to identify where possible misinterpretation occurs and which items do not appear to measure the construct of interest, while producing information about the structure of the rating scale and the degree to which

> each item contributes to the construct. Thus, it provides a mathematically sound alternative to traditional approaches to survey data analysis (p. 13).

Although the specifics of the Rasch model are beyond the scope of this paper, those interested in learning more about Rasch measurement should visit http://www.rasch.org/memos.htm for a wonderful compilation of measurement research. To exhibit the power and utility of Rasch measurement, a demonstration of just one of its powerful techniques, particularly the use of person and item maps will be provided. It should be noted that under CTT models and traditional statistical software packages, this technique cannot be performed.

Person and Item Maps

Person and item maps are incredibly useful and easy to interpret. These maps have the ability to place both persons and items on the same scale, thus allowing for meaningful comparisons. Figure 1 presents an item map produced from a previously published article by Bradley, Royal, Cunningham et al. (2008). The survey instrument sought to solicit graduate students and university faculty perceptions of what constitutes quality education research. Because the items were rather lengthy, they were coded simply by construct. For example, "M1", "M2", "M3", etc. would refer to items pertaining to methodological issues. Items "E1", "E2", "E3"… and "T1", "T2", T3", etc. would refer to items pertaining to ethical and theoretical issues, respectively. For the purposes of this demonstration, only a few select items will be highlighted to illustrate the interpretation of the map. This method is intended to allow readers of this manuscript to envision their data in the map, as opposed to focusing too much on the survey content provided.

Understanding the Map

First, it is helpful to understand the layout and design of the map. The numbered, vertical column on the left indicates logit measures. Before proceeding further it is important to briefly provide some background to explain logit measures. Logits are the measures

produced from raw scores when computed via the Rasch model. That is, the ordinal data that would appear as raw scores in a survey (i.e., ratings of 1, 2, 3, and 4) are converted to their natural logarithm, thus producing a measure that fits an interval scale. This conversion to truly interval data is at the heart of quality measurement and is one of the key distinctions of sound measurement. Once the raw score to logit conversion is complete, results are then interpretable. So, the numerical column on the left ranging from 4 to -3 indicates a ruler of logit scores. It should be noted that Rasch analysis produces logits for both persons and items, estimates which contain four decimal places and are quite exact. Person and item maps simply serve as a visualization of these findings. Therefore, precise logit measures would need to be found in other forms of Rasch output.

Second, notice the map is delineated into two halves. The left side of the map contains person measures and the right side contains item measures. Also, note that both person and item measures are placed along the same ruler, thus making it useful for easy and meaningful interpretation. Next, let us consider the marker line running vertically in the middle of the map. This line separates the two halves of the map. Notice, both sides of the map contain the symbols M, S, and T. These markers denote the mean (M), one standard deviation (S), and two standard deviations (T) for both persons and items. In this particular example, we can see that the person mean falls around 1.8 logits and the item mean falls around 0 logits. We can also see that two of the items “M12” and “M15” fall at approximately 3 logits and both items are beyond two standard deviations from the item mean.

The final step in interpreting the item map is understanding the hierarchy produced. It is possible to produce the hierarchy in various ascending and descending order with regard to person and item distributions. However, for this demonstration the default hierarchy generated from these maps will be illustrated. In this example, persons at the top of the map indicate they found it easier to endorse (or agree with) items than persons situated below

them on the map. Essentially, the idea is that persons at the top had the least difficulty endorsing items, while persons at the very bottom had the most difficulty endorsing items. Items can be interpreted in a similar manner. Items at the very top of the map were the most difficult to endorse, whereas items at the bottom of the map were the easiest to endorse.

Interpreting the Map

To provide some context to the example provided and to aid in the interpretation of results, survey items will be provided in full for items at the extreme ends of the item map. That is, item "M12" refers to the statement "High-quality research requires random sampling" and "M15" refers to the statement "High-quality research can be determined solely by examining the research methodology". Additionally, item "E1" refers to the statement "High-quality research abides by ethical standards" and item "E3" refers to the statement "High-quality research should protect the safety and welfare of participants".

Based on the map, items "M12" and "M15" fell at the top, which indicates they were the two most difficult items to endorse. If one imagines a horizontal line spanning across the entire map from this row of items one could see the proportion of persons who found these two items easy to endorse as well as the proportion of persons who found these items difficult to endorse. Here, persons who fell in the range of 3 to 4 logits easily endorse these items. The vast majority of persons, however, fell 3 logits and below, suggesting respondents had a more difficult time endorsing these items. Depending on how far the person measures are away from 3 logits essentially answers the question of how difficult it was for each of the various respondents to endorse these items. The further down the map the persons fell, the more difficulty they would have endorsing the two items at the top.

Also, notice the items at the bottom of the map, in particular, items "E3" and "E1". Following the same procedure as before, if one were to draw a line across these items through the person side of the map he or she would not find any persons at or below the measure. In

fact, one would have to go up to approximately -0.3 logits on the scale before the lower end of the person measures appeared. So, what does this mean? It means that items that fell at or below approximately -0.3 logits were rather easy for respondents to endorse.

To add a quality control element to this example, consider investigating the abbreviated items in the map. Items “M12” and “M15” read “High-quality research requires random sampling” and “High-quality research can be determined solely by examining the research methodology”. It makes sense that these items would be the most difficult to endorse, as these perspectives are likely to represent a minority perspective among faculty and graduate student researchers who research issues in education. Items “E1 and “E3” read “High-quality research abides by ethical standards” and “High-quality research should protect the safety and welfare of participants”. Because one would expect this sample frame to generally agree with these statements, it is safe to assume the hierarchy is ascending/descending in the direction consistent with this interpretation.

Software and Analytical Techniques

Data analysis for this demonstration was performed using Winsteps measurement software (Linacre, 2009). Traditional statistical analyses involve coding data, importing/uploading to a statistical software program, choosing appropriate statistical techniques, performing analyses, and so on. In many ways, Winsteps measurement software streamlines the process and actually makes it less likely to make an error. Similar to a traditional statistical analysis a data file must be prepared. However, measurement software requires a control file which contains commands for telling the program where to read the data, as well as what output to produce. Unlike statistical software packages such as SPSS, Minitab, etc. that require the researcher to choose a particular technique then perform the desired analyses, Winsteps measurement software practically produces all the output with one click of the button. This reduces the risk of the researcher choosing an inappropriate analysis

technique or possibly making a mistake. In other words, the researcher can be less concerned about choosing an appropriate statistical technique and more concerned about how to interpret results.

Like getting acquainted to any analytical software package, understanding programs like Winsteps will require some learning curve. Online manuals containing most everything one would need to know, as well as countless examples can be found at the software's homepage. For those interested in trying Winsteps measurement software, an evaluation/student version called MINISTEP is available free of charge at www.winsteps.com. The evaluation/student version allows analyses of records containing up to 75 persons and 25 items. This would be ideal for researchers wishing to survey a sample of less than 75 students/faculty/administrators, etc. with a questionnaire containing 25 survey items or less (excluding demographic questions). The full version has the capacity to analyze data sets containing up to 30,000 items and 10,000,000 persons.

Conclusion

The majority of survey research today (both published and unpublished) is lacking with regards to quality measurement. Reporting means and standard deviations based on ordinal measures is an inappropriate, yet widespread practice in the arena of higher education research. Utilizing measurement techniques to analyze data can correct many of the erroneous assumptions made by CTT models. This demonstration illustrated just one of the many useful measurement tools available to properly handle survey data.

It is the researcher's hope that assessment and institutional research practitioners will explore issues of measurement within their own research. It is important to understand that Rasch measurement is not intended to take the place of statistics, but rather to complement the use of statistics by serving as a prerequisite for data analysis. Utilizing a theoretically-sound and mathematically-just approach like Rasch measurement eliminates many

assumptions researchers often make regarding methodological issues. Therefore, once proper measurement takes place, appropriate statistical techniques can then be applied and the results will become more precise, and possibly more meaningful.

References

Becker, G. (2001). Controlling decremental and inflationary effects in reliability estimation resulting from violations of assumptions. *Psychological Reports, 89,* 403-424.

Bond, T. & Fox, C. (2001). *Applying the Rasch model: Fundamental measurement in the human sciences*. Mahwah, NJ; Lawrence Erlbaum Associates, Inc.

Bradley, K.D., & Sampson, S. (Spring, 2005). A case for using Rasch Rating Scale analysis to assess the quality of measurement in survey research. *The Respondent,* 12-13.

Bradley, K.D., Royal, K.D., Cunningham, J.D., Weber, J.A., Eli, J.A. (2008). What constitutes good educational research? A consideration of ethics, methods and theory. *Mid-Western Educational Researcher, 21*(1), 26-35.

Bunderson, C.V. (2000). *Design experiments, design science, and the philosophy of measured realism: Philosophical foundations of design experiments*. Paper presented at the Annual Meeting of the American Educational Research Association, New Orleans, LA. April 24-28.

Hambleton, R.K., Swaminathan, H., & Rogers, J.H. (1991). *Fundamentals of item response theory*. New York: Sage publications.

Linacre, J.M. (2009). Winsteps® (Version 3.68.0) [Computer Software]. Beaverton, Oregon: Winsteps.com.

Moulton, M. (2009). One ruler, many tests: A primer on test equating. EDS Publications. Available online at: http://www.eddata.com/resources/publications/EDS_APEC_Equating_Moulton.pdf

Spearman, C. (1904). “General intelligence” objectively determined and measured. *American Journal of Psychology, 15,* 201–293.

Stevens S. S. (1959). *Measurement, Psychophysics and Utility,* in C. W. Churchman & P. Ratoosh (Eds.), *Measurement: Definitions and Theories*. New York: John Wiley.

Figure 1. Map of survey items by difficulty to endorse