***Statistics Corner***

Questions and answers about language testing statistics:

# Likert items and scales of measurement?

James Dean Brown (University of Hawai‘i at Mānoa)

**Question:** Many people have asked me this seemingly simple question: Are “Likert-scale” questions on questionnaires nominal, ordinal, interval, or ratio scales?

**Answer:** In preparing to answer this seemingly easy question, I discovered that the answer is far from simple. To explain what I found, I will have to address the following sub-questions:

1. What are scales of measurement?
2. What does the literature say about Likert items and scales of measurement?
3. What does common sense tell us about Likert items and scales of measurement?

## *What are Scales of Measurement?*

Language researchers commonly describe the different ways they measure things numerically in terms of *scales of measurement*, which come in four flavors: nominal, ordinal, interval, or ratio scales. Each is useful in its own way for quantifying different aspects of language teaching and learning.

*Nominal scales* categorize. A nominal scale can be based on natural categories like gender (male or female) or artificial categories like proficiency (elementary, intermediate, or advanced proficiency groups). Nominal scales are also sometimes called *categorical scales*, or *dichotomous scales* (when there are only two categories).

*Ordinal scales* order or rank things. For instance, an item might ask students to rank ten types of classroom activities from most to least interesting (from 1 through 10). The most interesting activity would be first, followed by second, third, etc. (sensibly, ordinal scales are most often expressed as ordinal numbers). While the order is clear on such a scale, it is not clear what the distances are along the ordering. Thus the 1st activity might be much more interesting than the 2nd, but the 2nd activity might be only a little more interesting than the 3rd, and so forth. In short, ordinal scales show us the order, but not the distances between the rankings. Such ordinal scales are also sometimes called *ranked scales*.

*Interval scales* show the order of things, but with equal intervals between the points on the scale. Thus, the distance between scores of 50, 51, 52, 53 and so forth are all assumed to be the same all along the scale. Test scores are usually treated as interval scales in language research. Scales based on Likert items are also commonly treated as interval scales in our field.

*Ratio scales* differ from interval scales in that they have a zero value and points along the scale make sense as ratios. For example, a scale like age can be zero, and it makes sense to think of four years as twice as old as two years.

Researchers are often concerned with the differences among these scales of measurement because of their implications for making decisions about which statistical analyses to use appropriately for each. At times, they are discussed in only three categories: nominal, ordinal, and *continuous* (i.e., interval and ratio are collapsed into one category). [For more on scales of measurement, see Brown, 1988, pp. 20-24; 2001, pp. 17-18.]

## *What Does the Literature Say About Likert Items and Scales of Measurement?*

*Likert items* were first introduced by Rensis Likert (1932). The following is an example of three Likert (pronounced /lɪkərt/) items:

| Statements | Strongly disagree | Disagree | Neutral | Agree | Strongly agree |
|---|---|---|---|---|---|
| 1. I understand the difference between Likert items and Likert scales. | 1 | 2 | 3 | 4 | 5 |
| 2. I understand how to analyze Likert items. | 1 | 2 | 3 | 4 | 5 |
| 3. I like using Likert items. | 1 | 2 | 3 | 4 | 5 |

The example items have five options. They could equally well have 3, 4, 6, 7, or more options. [For more information on choosing the number of options and on how to write sound Likert items, see Brown, 2001, pp. 40-42, 44-54.]

When I first delved into the general literature on Likert items and scales of measurement, I found most articles were counter-intuitive and confusing. A number of articles argued or assumed that Likert items do not form an interval scale, but instead should be considered ordinal scales and should be analyzed accordingly (e.g., Coombs, 1960; Vigderhous, 1977; Jakobsson, 2004; Jamieson, 2004; Knapp, 1990; Kuzon, Urbanchek, & McCabe, 1996). Other articles proposed ways to get around this perceived ordinal/interval scale "problem" by proposing alternative Likert-like item formats such as the two-stage alternative offered by Albaum (1997) or the phrase completion alternative offered by Hodge and Gillespie (2003).

Despite all this discussion of the ordinal nature of Likert items and scales, most of the research based on Likert items and scales that I have seen in our field treats them as interval scales and analyzes them as such with descriptive statistics like means, standard deviations, etc. and inferential statistics like correlation coefficients, factor analysis, analysis of variance, etc. So you can see why I found the general literature counter-intuitive and confusing. For the most part, it says that we should treat Likert scales as ordinal scales, yet the research in my field consistently treats them as interval scales. How can these two positions be reconciled?

I believe that much of this ordinal/interval confusion arises from the fact that many authors use *Likert scale* to refer to both the Likert item type (items of the form shown above) and Likert scales (sums or averages of the results on sets of Likert items). For example, a questionnaire might have a total of 120 Likert items, divided into 12 Likert scales of 10 items each. If we carefully differentiate between Likert *items* and Likert *scales*, as I have done throughout this article, I think that much of the confusion will dissipate.

In addition, several papers have shown that *Likert scales* can indeed be analyzed effectively as interval scales (see for instance, Baggaley & Hull, 1983; Maurer & Pierce, 1998; and Vickers, 1999). Also, Allen and Seaman (1997, p. 2) support treating Likert scales as interval data with certain rather sensible provisos: "The "intervalness" here is an attribute of the data, not of the labels. Also, the scale item should be at least five and preferably seven categories. Another example of analyzing Likert scales as interval values is when the sets of Likert items can be combined to form indexes. However, there is a strong caveat to this approach: Most researchers insist such combinations of scales pass the Cronbach's alpha or the Kappa test of intercorrelation and validity. Also, the combination of scales to form an interval level index assumes this combination forms an underlying characteristic or variable."

In another vein, a number of authors have shown how Rasch analysis can be used to analyze and improve Likert scales as well as transform them into *true* interval scales. For more on this topic in the general literature, see Andrich (1978), Hagquist and Andrich (2004), Linacre (2002), Van Alphen, Halfens, Hasman, and Imbos (1994), and Waugh (2002); in the area of language research, see Sick (2006, 2009) or Weaver (2005, 2010).

## *What Does Common Sense Tell Us About Likert Items and Scales of Measurement?*

Because they confuse Likert items with Likert scales, many authors look at a single Likert item and conclude that the 1 2 3 4 and 5 options form an ordinal scale at best, and therefore data based on these scales must be analyzed as though they are ordinal. I have two responses to that form of "logic".

When your read that MacArthur graduated first in the West Point class of 1903, that means he was at the top of his class ahead of whoever was second, third, fourth, fifth, etc. What is it about any Likert item 1 2 3 4 5 (much less an Likert scale) that can be expressed in ordinal numbers? Is *strongly agree* fifth, ahead of *agree* at fourth, and *neutral* at third, *disagree* at 2nd, and *strongly disagree* at 1st? This doesn't make sense, even at the Likert item level, much less at the Likert scale level.

From a Likert scale perspective, even if we were to accept the erroneous idea that Likert items are ordinal, saying that the resulting data must be analyzed as though they too are ordinal is like saying that test items that are scored right or wrong are nominal so data based on them must be analyzed as though they are nominal. Test scores are usually based on nominal right/wrong items, yet the total scores are always treated as interval data in our field. If the single argument (that Likert item options are ordinal) is wrong, then the compound argument (that Likert scales are ordinal [sic] because Likert items are ordinal [sic]) is doubly wrong.

The one 100% sensible treatment I have found for this set of issues is found in Carifio and Perla (2007). On page 114, they list "the top ten myths and urban legends about 'Likert scales' and the counter argument and 'antidote' for each myth and urban legend." According to the authors, the following myths are ***WRONG***:

> *Myth 1*—There is no need to distinguish between a scale and response format; they are basically the same "thing" and what is true about one is true about the other.
> *Myth 2*—Scale items are independent and autonomous with no underlying conceptual, logical or empirical structure that brings them together and synthesizes them.
> *Myth 3*—Likert scales imply Likert response formats and vice versa as they are isomorphic.
> *Myth 4*—Likert scales cannot be differentiated into macro and micro conceptual structures.
> *Myth 5*—Likert scale items should be analyzed separately.
> *Myth 6*—Because Likert scales are ordinal-level scales, only non-parametric statistical tests should be used with them.
> *Myth 7*—Likert scales are empirical and mathematical tools with no underlying and deep meaning and structure.
> *Myth 8*—Likert response formats can without impunity be detached from the Likert Scale and its underlying conceptual and logical structure.
> *Myth 9*—The Likert response format is not a system or process for capturing and coding information the stimulus questions elicit about the underlying construct being measured.
> *Myth 10*—Little care, knowledge, insight and understanding is needed to construct or use a Likert scale.

Notice in particular Myths 1, 5, and 6, which are directly related to the topic of this column. For more details about these 10 myths, you should of course refer to the original article.

## *Conclusion*

The original question was: Are Likert-scale questions on questionnaires nominal, ordinal, interval, or ratio scales? My experience and my take on the literature lead me to believe that the following are true:

With regard to Likert *items* -

1. We must think about individual Likert items and Likert scales (made up of multiple items) in different ways.
2. Likert items represent an item format not a scale.
3. Whether Likert *items* are interval or ordinal is irrelevant in using Likert *scale* data, which can be taken to be interval.
4. If a researcher presents the means and standard deviations (interval scale statistics) for individual Likert items, he/she should also present the percent or frequency of people who selected each option (a nominal scale statistic) and let the reader decide how to interpret the results at the Likert-item level.
5. In any case, we should not rely too heavily on interpreting single items because single items are relatively unreliable.

With regard to Likert *scales* -

1. Likert scales are totals or averages of answers to multiple Likert items.
2. Likert scales contain multiple items and are therefore likely to be more reliable than single items.
3. Naturally, the reliability of Likert scales should be checked using Cronbach alpha or another appropriate reliability estimate.
4. Likert scales contain multiple items and can be taken to be interval scales so descriptive statistics can be applied, as well as correlational analyses, factor analyses, analysis of variance procedures, etc. (if all other design conditions and assumptions are met).

## References

Albaum, G. (1997). The Likert scale revisited: An alternate version. *Journal of the Market Research Society, 39*, 331-349.

Allen, E., & Seaman, C. A. (2007). Likert Scales and Data Analyses. *Quality Progress, 40*, 64-65.

Andrich, D. (1978). A rating formulation for ordered response categories. *Psychometrika, 43*, 561-573.

Baggaley, A., & Hull, A. (1983). The effect of nonlinear transformations on a Likert scale. *Evaluation & the Health Professions, 6*, 483-491.

Brown, J. D. (1988). *Understanding research in second language learning*. Cambridge: Cambridge University Press.

Brown, J. D. (2001). *Using surveys in language programs*. Cambridge: Cambridge University Press.

Carifio, J., & Perla, R. J. (2007). Ten common misunderstandings, misconceptions, persistent myths and urban legends about likert scales and likert response formats and their antidotes. *Journal of Social Sciences, 3*(3), 106-116.

Coombs, C. H. (1960). A theory of data. *Psychological Review, 67*, 143-159.

Hagquist, C., & Andrich, D. (2004). Is the Sense of Coherence-instrument applicable on adolescents? A latent trait analysis using Rasch-modelling. *Personality and Individual Differences, 36*, 955-968.

Hodge, D. R., & Gillespie, D. (2003). Phrase completions: An alternative to likert scales. *Social Work Research, 27*, 45-55.

Jakobsson, U. (2004). Statistical presentation and analysis of ordinal data in nursing research. *Scandinavian Journal of Caring Sciences, 18*, 437-440.

Jamieson, S. (2004). Likert scales: How to (ab)use them. *Medical Education, 38*, 1212-1218.

Knapp, T. R. (1990). Treating ordinal scales as interval scales: An attempt to resolve the controversy. *Nursing Research, 39*, 121-123.

Kuzon, W. M. Jr., Urbanchek, M. G., & McCabe, S. (1996). The seven deadly sins of statistical analysis. *Annals of Plastic Surgery, 37*, 265-272.

Likert, R. (1932), A Technique for the measurement of attitudes. *Archives of Psychology, 140*, 1-55.

Linacre, J. M. (2002). Optimizing rating scale category effectiveness. *Journal of Applied Measurement, 3*(1), 85-106.

Maurer, J., & Pierce, H. R. (1998). A comparison of Likert scale and traditional measures of self-efficacy. *Journal of Applied Psychology, 83*, 324-329.

Sick, J. (2006). *The learner's contribution: Individual differences in language learning in a Japanese high school*. Doctoral dissertation, Temple University, Philadelphia, PA.

Sick, J. (2009). Rasch measurement in language education part 3: The family of Rasch models. *Shiken: JALT Testing & Evaluation SIG Newsletter, 13*(1), 4-10.

Van Alphen, A., Halfens, R., Hasman, A., & Imbos, T. (1994). Likert or Rasch? Nothing is more applicable than a good theory. *Journal of Advanced Nursing, 20*, 196-201.

Vickers, A., 1999. Comparison of an ordinal and a continuous outcome measure of muscle soreness. *International Journal of Technology Assessment in Health Care, 15*, 709-716.

Vigderhous, G. (1977). The level of measurement and 'permissible' statistical analysis in social research. *Pacific Sociological Review, 20*(1), 61-72.

Waugh, R. F. (2002). Creating a scale to measure motivation to achieve academically: Linking attitudes and behaviours using Rasch measurement. *British Journal of Educational Psychology, 72*, 65–86

Weaver, C. (2005). Using the Rasch model to develop a measure of second language learners' willingness to communicate within a language classroom. *Journal of Applied Measurement, 6* (4), 396-415.

Weaver, C. (2010). *Japanese university students' willingness to use English with different interlocutors*. Doctoral dissertation, Temple University, Philadelphia, PA.

### *Where to Submit Questions:*

Please submit questions for this column to the following e-mail or snail-mail addresses: *brownj@hawaii.edu*. Your question can remain anonymous if you so desire.

JD Brown, Department of Second Language Studies
University of Hawai'i at Manoa, 1890 East-West Road
Honolulu, HI 96822 USA

**HTML**: http://jalt.org/test/bro_34.htm / **PDF**: http://jalt.org/test/PDF/Brown34.pdf