JOURNAL
OF
APPLIED
QUANTITATIVE
METHODS

RATING SCALE OPTIMIZATION IN SURVEY RESEARCH: AN APPLICATION OF THE RASCH RATING SCALE MODEL

Kenneth D ROYAL, PHD
American Board of Family Medicine

E-mail: kroyal@theabfm.org

Amanda ELLIS
University of Kentucky

Anysia ENSSLEN
University of Kentucky

Annie HOMAN
University of Kentucky

Abstract: Linacre (1997) describes rating scale optimization as "fine-tuning" to try to squeeze the last ounce of performance out of a test [or survey]". In the survey research arena, rating scale optimization often involves collapsing rating scale categories and performing additional iterative analyses of the data to ensure appropriate fit to the Rasch model. The purpose of this research is to 1) explore the literature as it pertains to rating scales in survey research, 2) discuss Rasch measurement and its applications in survey research, 3) conduct an iterative Rasch analysis of a sample survey dataset that demonstrates how collapsing rating scale categories can sometimes improve construct, communicative and structural validity and increase the reliability of the measures, and 4) discuss the implications of this technique.

Quality rating scales are essential for meaningful measurement in survey research. Because rating scales are the communication medium between the researcher and survey respondents, it is important that "communication validity" is evident in all survey research (Lopez, 1996). Communication validity is the extent to which the survey was developed in a manner that is unambiguous in language, terminology, and meaning for respondents. Further, it is also the extent to which respondents were able to clearly identify the ordered nature of the rating scale response options and accurately distinguish the difference between each category. However, establishing communication validity can be a bit tricky as survey research comes in many shapes and sizes. For example, rating scales may solicit degrees of frequency, agreement, importance, quality, likelihood, and a host of other measures. Selecting a response option is not so uniform either. This may involve selecting a single

response from a dichotomous scale (e.g., yes/no), a trichotomous scale (e.g, yes/maybe/no), or scales with varying ranges and response options (e.g., Strongly Agree, Agree, Disagree, Strongly Disagree). Regardless of the scale used there is the risk of using too few categories which can result in inaccurate findings due to "lumpy" data, or too many categories which can result in better accuracy but also more confusion for the respondent. The sample-dependent nature of the survey also makes choosing the best scale increasingly problematic. It is not uncommon to pilot test a survey on a group of individuals only to find the rating scale did not function the same way for the larger sample. Similarly, this may also happen when an existing instrument and rating scale is administered to a sample from a different population. In any instance, it would be irresponsible and/or naïve to say a one-size-fits-all solution is available for choosing the ideal rating scale given research purposes and respondent samples are so different. Despite this inevitability, Rasch measurement models can optimize rating scale functioning and provide researchers with more meaningful results.

Linacre (1997) describes rating scale optimization as "'fine-tuning'" to try to squeeze the last ounce of performance out of a test [or survey]". Consider the following example. A survey developer envisioned respondents would utilize all the response options along a 7-point continuum to respond to various items, but in actuality, respondents only utilized 5 categories. As Ben Wright, the most notable proponent of Rasch measurement in the world would say, to analyze the scale as though the respondents conceptualized more levels than were actually conceptualized is to deceive ourselves. Rating scales that utilize more response options than survey respondents actually use are ideal candidates for collapsing. However, survey researchers should be warned that this is not always the case. Wright and Linacre (1992) offer guidelines for collapsing rating scales. The authors suggest that any collapsing of categories should above all else, make sense. However, they warn that it is possible that collapsing some qualitative categories may create an artificial category which can have negative effects on validity. To help avoid these pitfalls, survey developers are advised to create a histogram that displays the frequency for which each rating scale category was utilized. From there, one can visually inspect the extent to which the rating scale was utilized and begin to envision which categories would make the most sense to collapse without creating an artificial category. This process is one component of a larger quality control process that evaluates the structural validity of the rating scale. More discussion of these processes as well as a demonstration will be provided in the methodology.

Although rating scale optimization techniques are often implemented by psychometricians and others who are expert in Rasch measurement analyses, few survey researchers in arenas outside the immediate Rasch measurement circles employ these techniques. The authors argue that perhaps survey researchers should consider Rasch measurement, as there can be a great deal of utility in this practice. The purpose of this research is to 1) explore the literature as it pertains to rating scales in survey research, 2) discuss Rasch measurement and its applications in survey research, 3) conduct an iterative Rasch analysis of survey data that demonstrates how collapsing rating scale categories can sometimes improve construct, structural and communicative validity and increase the reliability of the measures, and 4) discuss the implications of this technique.

Literature Regarding Rating Scales

Survey research is perhaps the most popular data collection technique in the social and behavioral sciences. Questionnaires can be conducted verbally, using paper and pencil, or by computer. Questionnaires can be open-ended in which respondents reply with their own words to questions, or they may be closed form in which individuals respond to questions or statements using a specific type of scale. Open-ended questionnaires provide less standardized responses, and researcher bias may influence how the responses are interpreted (Converse & Presser, 1986). The less definitive answers and lengthy responses that often result from open-ended surveys make for more complicated data analysis. Closed forms allow for more specific answers, call for less interpretation from researchers, and improve the ease of data collection. Closed forms include ranked items, check lists, and response scales (McMillan & Schumacher, 2010).

Response scales imply the researcher chooses questions or statements followed by a scale of potential responses which measure intensity of respondents' opinions or beliefs (Nardi, 2006). Rating scales include various types such as Thurstone scales (1928) which are used to measure attitude toward a particular construct by having respondents agree or disagree with statements equating with a predetermined level of favorability for that construct. Guttman scales (1944) also use a dichotomous response format but have statements arranged from least extreme to most extreme so that respondents will have a point of transition from affirmative to negative answers. Semantic differential scales (Osgood, Suci & Tannenbaum, 1957) use bipolar adjectives and offer respondents a range to indicate how their preference toward one or the other descriptor as it relates to a particular construct.

However, the most commonly used scale was created by Rensis Likert in 1932. When responding to an item on a Likert scale, respondents are asked to specify their level of agreement to a given statement (McMillan & Schumacher, 2010). The creation of questions or stems for a Likert scale is both an art and a science. To ensure that the statements get to the heart of the question and remain objective, non-leading, unambiguous, and relatively concise may be quite challenging. In addition to the stems, the response options need careful consideration in order to avoid certain pitfalls and increase the likelihood that the scale will function as intended.

Respondents rely on the labels of a response scale to create meaning for the scale points (Klockars & Yamagishi, 1988). Based upon the purpose of the survey, a response scale may include numerical labels, verbal labels, or both. Ultimately, when scales are verbally labeled for each point, the measurement of validity improves (Krosnick & Fabrigar, 1997). However, while verbal labels have been shown to be more reliable and valid than numerical labels, terms such as "often" and "sometimes" may be ambiguous and result in inaccurate responses (Jamieson, 2004). As the data collection tool is developed, it is essential that the appropriate use of verbal labels remain a focal point for designing a survey that will yield accurate data (Klockars & Yamagishi, 1988).

Stylistic elements, the visual attributes such as background colors, font, and spacing, are features of a response scale that are not necessarily essential, but help to give a survey or questionnaire its "look and feel" (Tourangeau, Mick, Couper, & Conrad, 2004). Respondents will use purely visual cues to aid them in their interpretation of response scale items. Examples include: respondents tend to assume a middle position means typical, something positioned left and top means first, things placed near one another imply they are

related, items placed in the upper portion of the scale are viewed as good or positive, and similar appearance is interpreted as close in meaning (Tourangeau, Couper, & Conrad, 2004). In addition, when the extreme end points of the scale are shaded in different hues, responses tend to shift to the higher end of the response scale than when both extreme ends are shaded in the same hue (Tourangeau, Couper, & Conrad, 2007).

Two characteristics demonstrated by survey respondents, *acquiescence* and *extreme responses,* may have an effect on how answers are provided (Billiet & McClendon, 2000; Cheung & Rensvold, 2000; Moors, 2003). The idea of *extreme response* refers to a respondent's tendency to choose items at the extreme ends of a response scale (Moors, 2008). Those who have the option of a mid-point do not necessarily answer the question or item in the same way that they would if they were forced to "choose a side" about the issue being explored (Bishop, 1987; Kalton, Roberts, & Holt, 1980). When deciding whether to include a mid-point in a response scale, it is also imperative to consider how the mid-point option may be interpreted by respondents. Respondents may use this option when the middle category accurately describes their neutral position regarding an item on a response scale. Another interpretation is that a mid-point may equate to "no opinion" (Maitland, 2009). This response may be an "easy out" for respondents who are unwilling or unable to express their opinion due to the cognitive encumbrance of a particular response scale item (Krosnick, 1991).

Researchers usually desire their respondents to make a definite choice rather than to choose a neutral or middle position. Therefore, it may be said that a response scale without a mid-point or neutral point would be preferred as long as it did not affect the validity or the reliability of the responses (Garland, 1991). It has been suggested that the inclusion of a mid-point leads to lower reliability for shorter response scales that have fewer items (Alwin, 2007).

Response options should allow respondents to both sufficiently and accurately discriminate between scale options. In other words, a range of choices allows for better classification, but too many choices makes precision problematic (DeVellis, 1991). The cognitive processes needed for respondents to generate answers could be a factor in whether or not the respondent is optimizing or satisficing when giving their response (Krosnick, 1999). Satisficing implies that respondents are not putting forth full effort to provide their most thought-out response. A large number of response items offer no empirical advantage over a small number, and experiments suggest that four to seven categories be used to optimize validity and to provide consistent and reliable participant responses (McKelvie, 1978; Weng, 2004; Lozano, Garicai-Cueto, and Muniz, 2008).

Rasch Measurement

Rasch modeling is already a popular method of choice in the survey research arena particularly in the health sciences, business and psychology, but it is also quickly becoming the norm for establishing quality measurement and valid instruments in all the social sciences. Rasch models are logistic, latent trait models of probability for monotonically increasing functions. Unlike statistical models that are developed based on data, Rasch measurement models are static models that are imposed upon data. Rasch models are invariant and assume the probability of a respondent agreeing with a particular item is a logistic function of the relative distance between the person and item location on a linear

continuum. Rasch models require unidimensional data and may be utilized in both dichotomous and polytomous scenarios.

With survey research, polytomous models are often employed. When a survey utilizes a rating scale that is consistent with regard to the number of response options (i.e., a 5-point rating scale for all items), the Rating Scale Model (Andrich, 1978) would be the appropriate model to apply. The formulae for the Rating Scale Model are presented below:

$$\ln (P_{nik}/P_{ni(k-1)}) = B_n - D_i - F_k$$

where,
P_{nik} is the probability that person *n* encountering item $_i$ is observed in category $_k$,
$P_{ni(k-1)}$ is the probability that the observation (or response) would be in category $_{k-1}$,
B_n is the "ability" (attitude, etc.) measure of person $_n$,
D_i is the "difficulty" measure of item $_i$,
F_k is the impediment to being observed in category $_k$ relative to category $_{k-1}$.

In situations where the rating scale varies from item to item (i.e., some items utilize a 5-point rating scale, others use a different 4-point scale), the Partial Credit Model (Masters, 1982) would be the appropriate model to apply. The formulae for the Partial Credit Model are presented below:

$$\ln (P_{nik}/P_{ni(k-1)}) = B_n - D_{ik}$$

Although the process of Rasch analysis is well documented in the literature (see Wright and Stone, 1979; Wright and Stone, 1999; Smith, Jr. & Smith, 2004; and Bond & Fox, 2007), it would be suffice to say that the analysis is largely concerned with the extent to which observed data match what is expected by the model.

Method

Instrument and Data

The data utilized in this study is from an academic misconduct study of 262 undergraduate business and economics students from a southern university. A 20 item instrument asked students to rate the extent to which they believed each item would affect the frequency of academic misconduct. The rating scale consisted of five categories: 1-Definitely would reduce academic misconduct; 2-Probably would reduce academic misconduct; 3-Would not affect misconduct; 4-Probably would increase academic misconduct; and, 5-Definitely would increase academic misconduct.

Rating Scale Diagnostics

Investigating rating scale diagnostics is useful as it demonstrates the extent to which respondents utilized each rating scale option (see Table 1). Here, counts and percents are provided to illustrate these points. Infit and outfit mean square statistics provide information about how well each category fits the Rating Scale Model. Based on these data, it is apparent that most respondents utilized ratings 1-3, with rating 4 utilized some (7% of the time), and rating 5 seldom utilized (3% of the time). Based on this information, one could establish a case that ratings 4 and 5 should be collapsed and re-analyzed. Also, because

the scale is balanced and contains a midpoint (response category 3), one might wish to make the scale a trichotomy by collapsing 1 and 2 into a single category, maintaining 3 as a midpoint, and collapsing 4 and 5 into a single category as well. Because it would make sense, at least on the surface, to consider these scenarios, an iterative analysis will be performed to determine which, if any, scenario provides the most meaningful information and improves the quality of measurement taking place with these data.

Table 1

Rating Scale Diagnostics

Rating Scale Category	Count	Percent	Infit Mnsq	Outfit Mnsq
1=Definitely would reduce academic misconduct	943	19	1.03	1.01
2=Probably would reduce academic misconduct	1637	33	.86	.89
3=Would not affect misconduct	1884	38	.91	.89
4=Probably would increase misconduct	326	7	1.00	1.01
5=Definitely would increase misconduct	140	3	1.27	1.34

Iterative Analyses

Three separate analyses were performed using Winsteps (Linacre, 2010) measurement software based on the rationale mentioned above. The first analysis investigated the quality of measurement and rating scale functioning based on the 12345 rating scale schema provided to survey participants. The second analysis investigated the same criteria for a rating scale which collapsed categories 4 and 5 as they were rarely utilized by survey respondents. The third analysis collapsed categories 1 and 2 as they were both categories that referred to reducing academic misconduct, collapsed categories 4 and 5 as they were both categories that referred to increasing academic misconduct, and category 3 remained unaltered as it provided the neutral category in the scale.

Probability Curves

Probability curves provide an excellent tool to visually view how well the rating scale is functioning. With a rating scale that is functioning well, a series of distinguishable hills should be present. Each hill should somewhat stand alone, as hills that tend to blend in with other hills indicate categories which raters may have a found difficult to endorse. Below are the probability curves for the three analyses performed in this study.

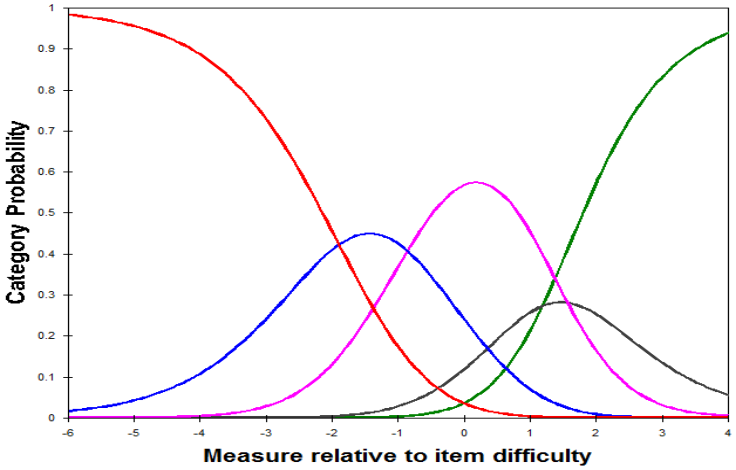

Figure 1. Rating Scale "12345"

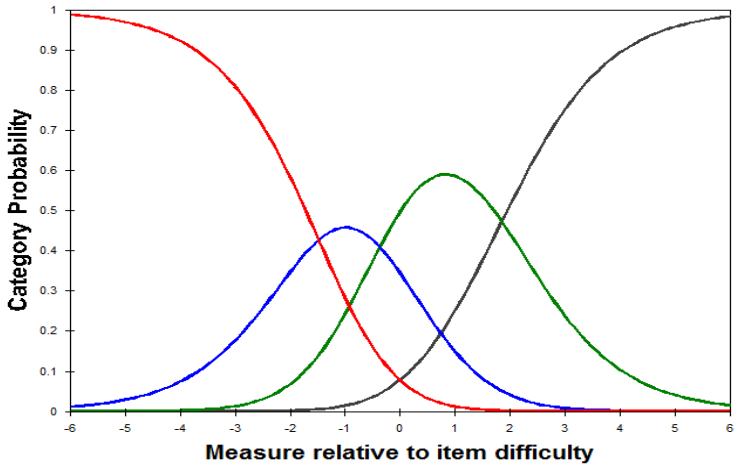

Figure 2. Rating Scale "Collapse 4,5"

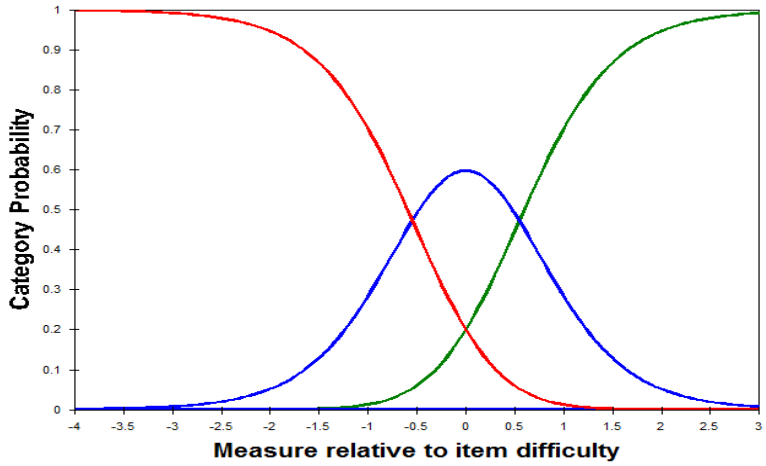

Figure 3. Rating Scale "Collapse 1,2; 4,5"

Figure 1 contains five hills, each indicating a rating scale response option (12345). Figure 2 illustrates four hills representing four rating scale categories (11244). Figure 3 illustrates three response categories (11355). Often probability curves will quickly indicate

rating scale problems as it relates to functioning, but here, all three figures illustrate hills that would be considered acceptable.

Results

Results of the three analyses are provided in Table 2. It appears collapsing categories 4 and 5 best optimized the use of the rating scale, as it improved separation and reliability measures and provided better data to model fit (improved validity) than the other two analyses. Leaving the scale unaltered by no means provided poor measurement, as the statistical indicators suggest the data fit quite well and the measures were sound. However, when the goal is maximize meaning, it is evident that collapsing categories 4 and 5 provided the most meaningful information.

Table 2

Reliability and Validity Measures

Rating	Separation		Reliability		Infit Mnsq		Outfit Mnsq	
Scale	Person	Item	Person	Item	Person	Item	Person	Item
12345	1.81	7.36	.77	.98	1.02	.98	1.00	1.00
Collapse 4,5	1.94	7.42	.79	.98	.99	1.00	1.01	1.00
Collapse 1,2; 4,5	1.50	7.17	.69	.98	1.03	1.00	.99	.98

Discussion and Conclusions

Initially, rating scale diagnostics indicated categories 4 and 5 of the rating scale were seldom utilized by survey respondents. For this reason, categories 4 and 5 were collapsed and the data were re-analyzed. Because categories 1 and 2 at the opposite end of the scale measure a related response "decreasing academic misconduct", it is within reason to test whether or not collapsing these categories would improve measurement as well. Data were re-analyzed on the trichotomous scale as well. Results indicate collapsing categories 4 and 5 improved measurement quality in this particular study. However, further collapsing categories 1 and 2 in addition to categories 4 and 5 negatively impacted the quality of measurement. In this particular instance data to model fit was not grossly affected by the additional collapsing, as fit indices suggested a negligible decrease in fit. However, reliability (one aspect of generalizability) dramatically decreased by about .10. An investigation of separation measures indicate the additional collapsing decreased the spread of the person measures, thus resulting in weaker reliability.

From the three analyses, results that indicate the most valid and reliable measures should be reported. In this case, the most meaningful measurement came from a rating scale that collapses categories 4 and 5. What does this mean for practice? Does it mean the researcher should collapse the rating scale to a 4-point scale for future administrations of the survey? Not necessarily. It would be perfectly acceptable to continue administering the survey in its present form. After all, there is no theoretical reason not to. Further, results indicate data fit the Rating Scale Model very well and the validity and reliability measures that were produced with the default 5-point scale were quite good. However, when the goal

is to "squeeze every ounce of performance out of a survey", as Mike Linacre aptly stated, one should use the strongest measures available to present results (2002).

Although rating scale optimization offers a number of potential benefits, not all survey researchers will embrace the practice. Many survey researchers are taught that rating scales should always be balanced, both in the data collection and reporting phases of research. Some researchers might contend balanced scale reporting is more aesthetically pleasing, or perhaps more true to form. Although the authors of the present study agree that balanced scales are generally well-advised, we contend that survey researchers should place a premium on obtaining meaningful results. That is, obtaining high-quality measures of validity and reliability are of the utmost concern. In some instances, following the "best practice" of balanced scales will impede our search for the most valid results and reliable measures possible. When this is the case, best practice reporting or conventions of aesthetics should be reconsidered.

Numerous studies have successfully optimized rating scales via collapsing categories. Although the present study demonstrates only a minor improvement of what is possible, several studies have shown significant advantages of this useful technique. Smith, Wakely, de Kruif, and Swartz (2003) collapsed a 10-point scale into a more meaningful 4-point scale, then re-administered the 4-point scale successfully in later administrations. Mark Stone (1998) was able to successfully collapse a 4-point Beck Depression Inventory into a dichotomy, as well as a 5-point survey focusing on fear. With regard to the fear dichotomy he points out, "Simple identification of fear is sufficient. Attempts to discriminate further are not useful" (p. 65). This is an excellent point, as sometimes the most useful and meaningful information resulting from survey analyses is simply whether or not a trait or phenomena is present or absent. Any attempts to dig further are futile in practice.

References

- Alwin, D.F. **Margins of Error: A Study of Reliability in Survey Measurement**. New York: John Wiley and Sons, Inc., 2007.
- Andrich, D. **A rating formulation for ordered response categories** *Psychometrika*, 43, 1978, pag. 561-73.
- Billiet, J.B., and McClendon, M.J. **Modeling Acquiescence in Measurement Models for Two Balanced Sets of Items**, *Structural Equation Modeling* 7, 2000, pag. 608-28.
- Bishop, G.F., **Experiments with the Middle Response Alternative in Survey Questions**, *The Public Opin. Q.* 51, 1987, pag. 220-32.
- Bond, T., Fox C., **Applying the Rasch model: Fundamental measurement in the human sciences**, *2nd edition*. Mahwah, NJ: Lawrence Erlbaum Associates, 2007.
- Cheung, G.W., Rensvold, R.B., **Assessing Extreme and Acquiescence Response Sets in Cross-Cultural Research Using Structural Equations Modeling**, *Journal of Cross-Cultural Psychology* 31, 2000, pag. 187-212.
- Converse, J.M., Presser S., **Survey Questions: Handcrafting the Standardized Questionnaire**, Quantitative Applications in the Social Sciences: Sage Publications, 1986

- DeVellis, R. **Scale development: Theory and applications** (2nd edition): Sage Publications, 1991.
- Garland, R. **The Mid-Point on a Rating Scale: Is It Desirable?** *Marketing Bulletin* 2 1991, pag 66-70.
- Jamieson, S. **Likert Scale: How to (Ab)Use Them,** *Medical Education* 38, 2004, pag. 1212-18.
- Kalton, G., Roberts J., Holt, D. **The Effects of Offering a Middle Response Option with Opinion Questions** *Statistician* 29, 1980, pag. 65-78.
- Klockars, A., Yamagishi, M., "**he Influence of Labels and Positions in Rating Scales,** *Journal of Educational Measurement* 25, no. 2, 1988, pag. 85-96.
- Krosnick, J. **Response Strategies for Coping with the Cognitive Demands of Attitude Measures in Surveys,** *Applied Cognitive Psychology* 5, 1991, pag. 201-19.
- Krosnick, J., **Survey Research**, *Annual Review of Psychology* 50, 1999, pag. 537-67.
- Krosnick, J., Leandre F., **Designing rating scales for effective measurement in surveys**, in L. Lyberg, P. Biemer, M. Collins, L. Decker, E. DeLeeuw, C. Dippo, N. Schwarz, and D. Trewin (Eds). *Survey measurement and process quality*: NY: John Wiley & Sons, 1997.
- Likert, R. **A technique for the measurement of attitudes**, *Archives of Psychology, 140,* 1932, pag. 1-55.
- Linacre, J. M. **Guidelines for Rating Scales** MESA Research Note #2, 1997, Available at: http://www.rasch.org/rn2.htm.
- Linacre, J. M., **Understanding Rasch measurement: Optimizing rating scale category effectiveness** *Journal of Applied Measurement, 3*(1), 2002, pag. 85-106.
- Linacre, J. M., **Winsteps** (Version 3.69.1) [Computer Software]. Beaverton, Oregon: Winsteps.com. 2010, Available from http://www.winsteps.com/.
- Lopez, W. **Communication Validity and Rating Scales** *Rasch Measurement Transactions, 10*(1), 1996, pag. 482-483.
- Lozano, L., Garcia-Cueto, E., Muniz, J., **Effect of the Number of Response Categories on the Reliability and Validity of Rating Scales,** *Methodology* 4, no. 2, 2008, pag. 73-79.
- Maitland, A., **How Many Scale Points Should I Include for Attitudinal Questions?** http://surveypractice.org/2009/06/29/scale-points/.
- Masters, G.N., **A Rasch model for partial credit scoring**, *Psychometrika*, *47*, 1982, pag. 149-174.
- McKelvie, S., **Graphic Rating Scales - How Many Categories?**, *British Journal of Psychology* 69, 1978, pag. 185-202.
- McMillan, J.H., Schumacher, S., **Research in Education: Evidence-Based Inquiry**, 7th ed: Pearson, 2010.
- Moors, G., **Diagnosing Response Style Behavior by Means of a Latent-Class Factor Approach: Sociodemographic Correlates of Gender Role Attitudes and Perceptions of Ethnic Discrimination Reexamined**, *Qual Quant* 37, 2003, pag. 277-302.
- Moors, G., **Exploring the Effect of a Middle Response Category on Response Style in Attitude Measurement**, *Qual Quant* 42, 2008, pag. 779-94.

- Nardi, P.M. **Doing Survey Research: A Guide to Quantitative Methods**, 2nd ed, Pearson Education, Inc., 2006.
- Osgood, Charles, George Suci, and Percy Tannenbaum. *The measurement of meaning*. Urbana, IL: University of Illinois Press, 1957.
- Smith, E. V., Wakely, M. B., Kruif, R. E. de, Swartz, C. W., **Optimizing rating scales for self-efficacy (and other) research**, *Educational and Psychological Measurement*, 63, 2003, pag. 369-391.
- Stone, M. H., **Rating scale categories. Dichotomy, Double Dichotomy, and the Number Two**, *Popular Measurement,* 1(1), 1998, pag 61-65
- Thurstone, L. L., **Attitudes can be measured**, *American Journal of Sociology,* 33, 1928, pag 529-54.
- Tourangeau, R., Couper, M.P., Conrad, F., **Color, Labels, and Interpretive Heuristics for Response Scales**, *Public Opinion Quarterly* 71, no. 1, 2007, pag. 91-112.
- Tourangeau, R., Mick, P., Couper, M.P., Conrad, F.G., **Spacing, Position, and Order: Interpretive Heuristics for Visual Features of Survey Questions**, *Public Opinion Quarterly* 68, 2004, pag. 368-93.
- Weng, L., **Impact of the Number of Response Categories and Anchor-Labels on Coefficient Alpha and Test-Retest Reliability**, *Educational and Psychological Measurement* 64, no. 6, 2004, pag. 956-72.
- Wright, B. D., Linacre, J., **Combining and Splitting Categories**, *Rasch Measurement Transactions, 6*(3), 1992, pag. 233-235.