Educational and Psychological
Measurement
70(5) 796–807
© The Author(s) 2010
Reprints and permission:
sagepub.com/journalsPermissions.nav
DOI: 10.1177/0013164410366694
http://epm.sagepub.com

Measuring the Mathematical Attitudes of Elementary Students: The Effects of a 4-Point or 5-Point Likert-Type Scale

Jill L. Adelson[1] and D. Betsy McCoach[2]

Abstract

The purpose of this study was to compare how students in Grades 3 to 6 respond to a mathematics attitudes instrument with a 4-point Likert-type scale compared with one with an additional neutral point (a 5-point Likert-type scale). The 606 participating students from six elementary and middle schools randomly received either the 4-point or 5-point format of the Math and Me Survey. Regardless of whether a neutral midpoint was offered or not, the structure of the instrument was virtually the same, with equal intercepts, means, variances and covariances, pattern coefficients, and nearly all residuals. The 5-point scale is preferred with this population because with this format the reliability estimate for the Mathematical Self-Perceptions subscale was higher ($p = .049$), and the pattern coefficients were stronger. Additionally, this format provided less model misfit than the 4-point format. Based on these findings, the authors recommend administration of the Math and Me Survey in the 5-point format. These findings also indicate that despite what some educators and educational experts believe, children in Grades 3 to 6 are capable of discriminating among five response options and do not tend toward the neutral point more so than with a 4-point scale.

Keywords

Likert-type scale, attitudes instrument, survey design, elementary students, mathematics attitudes

[1]University of Louisville, Louisville, KY, USA
[2]University of Connecticut, Storrs, CT, USA

Corresponding Author:
Jill L. Adelson, Department of Educational and Counseling Psychology, University of Louisville, Louisville, KY 40292, USA
Email: jill.adelson@louisville.edu

Educational researchers often are interested in assessing the attitudes of students, parents, teachers, and administrators. One of the most commonly used types of attitude scales in social science research is the Likert (1932) summated rating scale. Instruments designed with a Likert-type scale present respondents with statements or questions, each of which they respond to in terms of an agreement or preference continuum. This continuum typically ranges between extremes, such as disagree–agree or dislike–like.

For many decades, instrument developers have debated the issue of whether they should present an even or odd number of categories (Gable & Wolf, 1993; Ory & Wise, 1981). This issue affects the presentation of a neutral point as well as the optimal number of categories. Some have argued that including a middle category allows respondents to indicate a neutral response and be *more* discriminating in their response, making the scale scores more reliable and the scale more preferred by subjects (Cronbach, 1950; Gable & Wolf, 1993; Ory & Wise, 1991). On the other hand, others have expressed concern that with a middle category respondents will be *less* discriminating and declare themselves neutral more often, while omitting the neutral point forces respondents to be more thoughtful, resulting in more precise ratings. In fact, Garland (1991) indicated that a scale without a neutral middle is preferable because respondents are forced to make a definite choice, but he did acknowledge that this preference is overridden if the lack of a midpoint affects the validity or reliability of the responses. Busch (1993) and Reid (1990) also expressed concern about using a neutral midpoint as “neutrality can lead to indecisive data, particularly among those ethnic groups whose cultures value indirect responses” (Busch, 1993, p. 735).

Another concern with the optimal number of points or inclusion of a neutral point is the age of the subjects. According to Bourke and Frampton (1992), younger respondents might be more comfortable with fewer response categories, which would increase the reliability of scores because the respondents would respond more consistently. Educational experts have expressed concern that surveys should not present younger respondents with “too many” response categories. During the development of a mathematics attitudes instrument for elementary students in Grades 3 to 6 (Adelson & McCoach, 2009), educational experts served as content validators and weighed in on the issue of the number of scale points that should be presented to elementary students. Of the 14 experts, 10 recommended using a 4-point scale with 7 of those 10 feeling strongly or very strongly about their recommendation. The majority of the experts commented that students in this age range need to be forced to make a decision one way or the other and that too many options would interfere with their ability to discriminate the points. Thus, they suggested using a scale with 4 points because it does not have the neutral point and has fewer points in general than a 5-point scale. However, these suggestions were based on their beliefs about children and not on instrument design experience or empirical evidence. Although some researchers have examined the optimal number of response categories (e.g., Aiken, 1983; Bourke & Frampton, 1992; Champney & Marshall, 1939; Chang, 1994; Johnson, Smith, & Tucker, 1982; Ko, 1994; Komorita & Graham, 1965; Masters, 1974; Matell & Jacoby, 1971; Oaster, 1989; Preston & Colman, 2000; Weng, 2004; Wong, Chuen, & Fung, 1993), this research has not been conducted on elementary-aged children, despite the belief that children are not capable of the finer discrimination required by more points (five rather than four) and will tend toward a middle category if given the option.

Literature on the Optimal Number of Response Categories

Since the Likert format was first introduced, many researchers have tried to determine the optimal number of response categories by examining the relationship between the number of response categories and internal consistency, but the findings have been conflicting, with some indicating no effect of number of response categories on coefficient alpha (e.g., Aiken, 1983; Wong et al., 1993) and others indicating an effect but recommending different number of response options from 2 or 3 categories (Matell & Jacoby, 1971; Johnson et al., 1982) up to as many as 18 categories (Champney & Marshall, 1939). These results have not consistently favored an even or odd number of points, with some researchers recommending a range of response categories that includes both even and odd numbers, such as Ko's (1994) and Oaster's (1989) recommendation for 6 or 7 response categories and Preston and Colman's (2000) recommendation for 7 to 10 categories. Despite decades of research, the optimal number of response categories for Likert rating scales is still undetermined (Preston & Colman, 2000). Hypothesized explanations for the variability in the effects of the number of response categories on internal consistency reliability include the scale's item homogeneity (Bourke & Frampton, 1992; Komorita & Graham, 1965; Weng, 2004) and the variability of opinions (Masters, 1974; Weng, 2004). (For more information on effects on internal consistency reliability, see Henson, 2001).

Throughout the majority of the studies on the optimal number of response options, which have no general consensus, the researchers relied on measures of reliability, particularly internal consistency. As Chang (1994) noted, a problem with reliability studies is that "none of the studies used a model-fitting approach to determine which scale better fit the data. Simply comparing two reliability coefficients, as all existing studies have done, ignores other measurement considerations" (p. 205). Whereas the halo effect and response set, among other possible systematic errors due to the number of response categories, would increase internal consistency reliability artificially by altering homogeneity of the items of the scale, they would not do the same for validity coefficients; "therefore, validity is a better criterion than reliability in evaluating the optimal number of scale points" (p. 206). This is supported by Cronbach (1950), who acknowledged that "there is no merit in enhancing test reliability unless validity is enhanced at least proportionally" (p. 22). Therefore, researchers should consider *both* reliability and validity when studying the optimal number of response options, as Chang (1994) did when using confirmatory factor analyses of a multitrait–multimethod covariance matrix to compare 4-point and 6-point scales and as Bourke and Frampton (1992) did when comparing the stability of the factor structure of an instrument over 4-point, 5-point, 6-point, and reverse 4-point formats.

Nearly all these studies were conducted on adults, or at the youngest, adolescence. Halpin, Halpin, and Arbet (1994) did compare the internal consistency reliability when they altered the number and type of item–response choices to students with a mean age of approximately 12 years. They found that the estimates of Cronbach coefficient alpha increased with a 4-point Likert-type scale instead of a true/false format. This research has not been duplicated or extended beyond the comparison of a Likert-type scale to a dichotomous format or to include a neutral point. However, researchers such as Scott (1997) have encouraged instrument developers to take into account young respondent's cognitive and social capacities as "one can speculate that children and juveniles answer survey questions from a different cognitive and communicative perspective compared to adults" (Fuchs, 2002, p. 1). The concern expressed by the educational experts who served as content validators in the Adelson and McCoach (2009) study echoes Symonds's (1924) explanation of why the number of response categories may affect test reliability. When children, or any respondent, respond to a scale that requires finer discrimination than they typically can accomplish easily (too many response categories), this adds measurement error to their test scores as they cannot distinguish reliably between adjacent categories. However, with too few response categories, the scale may not elicit information on individual differences and would have less variability. Both these cases would lead to lower reliabilities. Thus, it is important for us to determine whether elementary-aged children can discriminate between categories on a 5-point Likert-type scale to the same degree they can on a 4-point Likert-type scale and how these formats affect the structure and reliability of the scales.

Method

Participants

A teacher, principal, or math coordinator at six elementary and middle schools from across the United States agreed to have students in their school in Grades 3 through 6 participate in the survey. Only students with parental permission completed the survey. Through coordination with their teacher, 606 students participated in the study. The researchers interweaved copies of the 4-point and 5-point formats prior to distribution to schools and classrooms, allowing random assignment of the number of points on the scale. Half of the students ($n = 304$) randomly received the survey with a 4-point Likert-type scale, and the other half ($n = 302$) randomly received the survey with a 5-point Likert-type scale. The percentage of male students was similar for each type of scale (45% for the 4-point scale and 41% for the 5-point scale) as were the percentages of various ethnicities (with about two thirds of students completing either scale being White). For the 4-point scale, each of the four grade levels made up 20% to 28% of the sample. For the 5-point scale, there were more fourth graders (35% of the sample) and fewer third graders (15% of the sample).

Instrument

Teachers administered the Math and Me Survey (Adelson, 2006) in both formats at the same time. The Math and Me Survey was developed and piloted for students in Grades

3 through 6. Previous exploratory and confirmatory factor analyses have resulted in identification of two factors, Enjoyment of Mathematics and Mathematical Self-Perceptions, which have a moderate correlation (.464). The Enjoyment of Mathematics subscale, which contains 10 items, measures to what degree a student takes pleasure in learning and doing mathematics. The Mathematical Self-Perceptions subscale, which contains 8 items, is a measure of a student's perception of his/her ability to learn and perform well in mathematics. The reliability coefficients of the pilot sample data for scores on both subscales, .920 (95% confidence interval [CI] = .908, .931; Zuo, 2007) for Enjoyment of Mathematics, and .880 (95% CI = .862, .897) for Mathematical Self-Perceptions, indicated a high degree of internal consistency, or domain homogeneity, among the subscale items.

This study involved the same survey formatted two ways. All the items were presented in the same order, but the Likert-type scale varied. On the surveys with a 4-point Likert-type scale, students responded to each item by circling one of the following: SD (*strongly disagree*), D (*disagree*), A (*agree*), or SA (*strongly agree*). For the 5-point Likert-type scale, students responded in a similar manner with the additional category of N (*neither agree nor disagree*) situated between *disagree* and *agree*.

So that responses to the two survey formats could be compared, it was necessary to code the response categories so that they spanned the same range, making mean, standard deviations, and pattern coefficient comparisons meaningful. For the 4-point Likert-type scale, the coding for the responses was –2 for *strongly disagree*, –.667 for *disagree*, .667 for *agree*, and 2 for *strongly agree*. For the 5-point Likert-type scale, the coding for the responses was –2 for *strongly disagree*, –1 for *disagree*, 0 for *neither agree nor disagree*, 1 for *agree*, and 2 for *strongly agree*.

Analyses and Results

Using confirmatory factor analysis in Amos 6.0, we tested the two-factor attitudes toward mathematics structure originally proposed by Adelson and McCoach (2009) independently for each of the survey formats. (Note that Amos ignores the ordinal nature of the data and considers the data to be continuous. However, we could not complete the analyses while accounting for the ordinal nature of the data because the thresholds naturally would be different for the two scales due to the difference in either intervals or range as it is impossible to constrain both.) The two latent variables were the two factors, Enjoyment of Mathematics and Mathematical Self-Perceptions. We specified a model in which each item was an indicator of only one factor, the one originally proposed to represent that item, and we permitted the two factors to be correlated. Because not every student responded to every item, we used maximum likelihood estimation to handle missing data, and we estimated the means and intercepts.

Examination of the path coefficients in the models revealed that the unstandardized regression weights were all statistically significant. The standardized regression weights, which represent the standardized path coefficients to the factors, were all in the desired range between .3 and .9 for both subscales. The path coefficient estimates for all but

Table 1. Fit Indices for Confirmatory Factor Analysis on 4-Point Scale and 5-Point Scale Separately

Index	4-Point Scale	5-Point Scale
χ^2	372.04	307.22
df	134	134
p	<.001	<.001
χ^2/df ratio	2.78	2.29
CFI	.92	.95
TLI	.90	.94
RMSEA (90% CI)	.077 (.067, .086)	.066 (.056, .075)

Note: *df* = degree of freedom; CFI = comparative fit index; TLI = Tucker–Lewis index; RMSEA = root mean square error of approximation; CI = confidence interval.

one item, "I do math problems on my own 'just for fun'" on the Enjoyment of Mathematics factor, were higher on the 5-point scale than on the 4-point scale. The greatest differences in path coefficient estimates were for "I enjoy playing math games" (difference of .11), "Math is very hard for me" (difference of .10), and "I can tell if my answers in math make sense" (difference of .90). Table 1 indicates the model fit indices for both formats. Using common rules of thumb for interpretation of model fit, the two-factor model is a fairly good fit, with the 5-point scale having less model misfit than the 4-point scale.

Having established reasonable fit of the structure for the instrument in both formats, although fit indices tended to favor the 5-point scale, we next examined the descriptive statistics for each format. For the total score on each of the two factors, there was no statistically significant difference in the means for the two survey formats, Wilks's $\Lambda = 1.00$, $F(2, 539) = 1.37$, $p = .25$. Similarly, regardless of survey format (4-point or 5-point scale), students did not exhibit statistically significant differences in their means on any item, using a conservative alpha to control for family-wise Type I error rates. Although one might expect more variance with the 5-point scale due to less restriction of range, the only item that exhibited statistically significantly unequal variances was "I am really good at math," with students being measured with the 5-point scale having a higher standard deviation. Interestingly, on nearly every item, the mean score for the 5-point scale was no closer to zero (the neutral point) than the mean score for the 4-point scale was, indicating that students were not more likely to choose *neither agree nor disagree* when given the option, contrary to what the content validators in the Adelson and McCoach (2009) study hypothesized would happen. The responses on the 5-point scale had more skew than the responses on the 4-point scale. About half of the items were more kurtotic on the 5-point scale, with the others being more kurtotic on the 4-point scale. Every item had a statistically significant Kolmogorov–Smirnov Z ($p < .001$), indicating that the cumulative distribution functions for the two formats are different in shape or location for every item. For all practical purposes though, the skew is not more visibly noticeable with the 5-point scale.

We compared the interitem correlation coefficients using Fischer's *r* to *z* transformations (Preacher, 2002). Like the means and standard deviations, the interitem correlations for both factors were similar for the two formats. In fact, despite most items seeming to have higher interitem correlations for the 5-point scale than for the 4-point scale (except for the item "I do math problems on my own 'just for fun'" on the Enjoyment of Mathematics subscale), no interitem correlations were statistically significantly different for the two formats (using a conservative significance level of .001 to control for family-wise Type I error rates).

Both subscales resulted in internally consistent scores for the sample. The reliability estimate for the Enjoyment of Mathematics subscale was similar (p = .413) for both formats; the 5-point scale had a coefficient α of .937 (95% CI = .926, .947), and the 4-point scale had a coefficient α of .928 (95% CI = .915, .940). (CIs and comparison of reliability estimates were conducted using SPSS; Fan & Thompson, 2001). Although acceptable for both formats, the reliability estimate for the Mathematical Self-Perceptions subscale for the 5-point scale (.911; 95% CI = .894, .926) was statistically significantly higher (p = .049) than it was for the 4-point scale (.878; 95% CI = .855, .898).

Next, we compared the structure of the two formats using multiple group confirmatory factor analysis. We maintained the same structure as when we analyzed the formats independently, with the two latent variables, paths leading to each item from the factor originally proposed to represent that item, and correlated factors. Once again, we estimated a means and intercept model because we had missing data. The two groups were those measured by the 4-point format and those measured by the 5-point format. First, we fit the proposed model to the data for each sample separately with only the intercepts constrained to be equal across groups. This is the minimal constraints allowable for a means and intercepts model (Arbuckle, 2005). This model served as the baseline model. Subsequently, in a hierarchical fashion, we placed more stringent constraints on the model by specifying the parameters of interest to be constrained across the group. Our constraints progressed in this fashion: intercepts, pattern coefficients, covariances and variances, means, and residuals. We then examined the model using the fit statistics for the more restrictive and the prior less restrictive model to determine whether the model and the individual parameter estimates are invariant across the two samples (representing the survey formats in this case). Once we found a more restrictive model that did not provide better model fit than the previous less restrictive model, we unconstrained partial parameters at that level, starting with the parameter with the greatest difference between groups and progressing to the next until the more restrictive model did provide equally good fit.

As shown in Table 2, each constrained model provided better model fit than the less restrictive model before it except the model in which all residuals were constrained to be equal across the two groups. This indicates that whether students completed the instrument in the 4-point or 5-point format that the instrument structure was nearly identical. Both formats had similar pattern coefficients, covariances and variances, and means. With the exception of three items from the Enjoyment of Mathematics factor—"I do math problems on my own 'just for fun,'" "I enjoy playing math games," and

Table 2. Multiple Group Comparisons (Invariance Tests) Across the Two Groups (Students Taking the 4-Point and 5-Point Formats)

Hypothesis Description	χ^2	*df*	χ^2/df	Models Compared	$\Delta\chi^2$	Δdf	*p*	CFI	TLI	RMSEA (90% CI)
1. Intercepts	706.14	284	2.49	—	—	—	<.001	.94	.92	.050 (.045, .054)
2. Model 1 + pattern coefficients	718.28	300	2.39	1 and 2	12.14	16	.73	.94	.93	.048 (.044, .053)
3. Model 2 + covariances and variances	720.72	303	2.38	2 and 3	2.45	3	.49	.94	.93	.048 (.043, .052)
4. Model 3 + means	723.96	305	2.37	3 and 4	3.24	2	.20	.94	.93	.048 (.043, .052)
5. Model 4 + residuals	771.45	323	2.39	4 and 5	47.48	18	<.001	.93	.93	.048 (.044, .052)
6. Model 5, without constraining E8, E4, S4, or E10	744.68	318	2.34	4 and 6	20.71	13	.08	.94	.93	.047 (.043, .052)

Note: *df* = degree of freedom; CFI = comparative fit index; TLI = Tucker–Lewis index; RMSEA = root mean square error of approximation; CI = confidence interval.

"Solving math problems is fun"—and one item from the Mathematical Self-Perceptions factor—"Math is very hard for me"—the residuals also were similar for both formats.

Discussion

The purpose of this study was to compare how students in the intermediate elementary grades (Grades 3-6) respond to two different formats of a mathematics attitudes scale—one with a 4-point Likert-type scale and one with an additional neutral point (a 5-point Likert-type scale). Educators have raised the question of children's ability to discriminate with five categories and have hypothesized that children would be more likely to gravitate toward the middle category if given the option (Adelson & McCoach, 2009), but prior to this study, these claims had not been examined empirically.

Despite the concerns expressed by educators, students responded in a similar manner, regardless of whether they responded to the 4-point or 5-point format. With the exception of four residuals, the two survey formats exhibited similar structure, including intercepts, pattern coefficients, covariances and variances, and means. The stability of the factor structure of the instrument over the two different formats is a measure of construct validity (Bourke & Frampton, 1992), indicating that the number of response categories (four or five) does not affect the construct being measured.

Although the two survey formats have similar structure as well as item means and variability, other evidence suggests that the 5-point format is favorable with this population. First, for all but one item, the standardized regression weights were stronger for the 5-point scale, and the 5-point scale exhibited less model misfit based on the χ^2/df ratio, the comparative fit index and Tucker–Lewis index, and the root mean square error of approximation estimate. Whereas the scores on the Enjoyment of Mathematics subscale exhibited similar internal consistency reliability for both formats, the Mathematical Self-Perceptions subscale exhibited statistically significantly higher reliability when measured with a 5-point scale. These results suggest that children can use a 5-point scale effectively and that 5-point scales outperform 4-point scales in terms of psychometric properties when used with this population. Thus, based on this research, the 5-point rating scale, which includes a neutral midpoint, seems to be appropriate even for elementary-aged children.

Limitations and Future Research

The primary limitation of this research centers on the potential generalizability of results. Halpin et al. (1994) suggested that the content measured in a scale greatly affects the optimal number of scale points. This study only examined the structure, internal consistency, model fit, and descriptive statistics for a scale measuring attitudes toward mathematics. Although the population was diverse, including students in Grades 3 through 6 from numerous states throughout the United States, the results found with this particular survey may not be generalizable to other surveys with this population. Future research should examine whether similar results are found using other surveys with children of this age.

Additionally, the only type of reliability that this study examined was internal consistency reliability. Students did not complete the survey a second time to examine test–retest reliability. Therefore, we cannot make any inferences regarding the stability of children's responses and how the number of response options affects response stability. Similarly, this study did not investigate the effects of the number of response options on discriminant or criterion validity. Future studies using the Math and Me Survey should investigate test–retest reliability and might include another measure of mathematical attitudes to explore how the number of response options affects other types of validity.

Finally, this study was limited to 4-point and 5-point scales. Further research must be conducted to explore how children respond to the survey with additional response options. It may be that students can discriminate with six or seven response options just as well and that model misfit might be further reduced. This might help address the question of whether the 5-point scale psychometrically outperformed the 4-point scale due to the additional response category or because of the addition of a neutral midpoint. Regardless of the need for future research examining the optimal number of response categories, due to the high reliability coefficients for both factors, researchers may confidently use the Math and Me Survey in the 5-point format to measure students' enjoyment of mathematics and mathematical self-perceptions.

Authors' Note

Additional tables providing standardized regression weights, comparison of descriptive statistics, assessment of normality, and comparison of item correlations are available from the authors on request. An earlier version of this article was presented at the annual meeting of the American Educational Research Association, San Diego, CA, April 2009.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the authorship and/or publication of this article.

Funding

The author(s) received no financial support for the research and/or authorship of this article.

References

Adelson, J. L. (2006). *Math and Me Survey*. Unpublished instrument.

Adelson, J. L., & McCoach, D. B. (2009). *Development and psychometric properties of the Math and Me Survey: Measuring third through sixth graders' attitudes towards mathematics*. Manuscript in preparation.

Aiken, L. R. (1983). Number of response categories and statistics on a teacher rating scale. *Educational and Psychological Measurement*, *43*, 397-401.

Arbuckle, J. L. (2005). *Amos 6.0 user's guide*. Spring House, PA: Amos Development.

Bourke, S., & Frampton, J. (1992, November). *Assessing the quality of school life: Some technical considerations*. Paper presented at the annual conference of the Australian Association for Research in Education, Deakin University, Geelong, Australia.

Busch, M. (1993). Using Likert scales in L2 research: A researcher comments. *TESOL Quarterly*, *27*, 733-726.

Champney, H., & Marshall, H. (1939). Optimal refinement of the rating scale. *Journal of Applied Psychology*, *23*, 323-331.

Chang, L. (1994). A psychometric evaluation of 4-point and 6-point Likert-type scales in relation to reliability and validity. *Applied Psychological Measurement*, *18*, 205-215.

Cronbach, L. J. (1950). Further evidence on response sets and test design. *Educational and Psychological Measurement*, *10*, 3-31.

Fan, X., & Thompson, B. (2001). Confidence intervals about score reliability coefficients, please: An EPM guideline editorial. *Educational and Psychological Measurement*, *61*, 517-531.

Fuchs, M. (2002, August). *Children and juveniles as respondents: Experiments on question order, response order and scale effects*. Paper presented at the International Conference on Improving Statistics, Copenhagen, Denmark.

Gable, R. K., & Wolf, M. B. (1993). *Instrument development in the affective domain* (2nd ed.). Boston, MA: Kuwer Academic.

Garland, R. (1991). The mid-point on a Likert rating scale: Is it desirable? *Marketing Bulletin*, *2*, 66-70.

Halpin, G., Halpin, G., & Arbet, S. (1994). Effects of number and type of response choices on internal consistency reliability. *Perceptual and Motor Skills*, *79*, 928-930.

Henson, R. K. (2001). Understanding internal consistency reliability estimates: A conceptual primer on coefficient alpha. *Measurement and Evaluation in Counseling and Development*, *34*, 177-189.

Johnson, S. M., Smith, P. C., & Tucker, S. M. (1982). Response format of the Job Descriptive Index: Assessment of reliability and validity by the multitrait-multimethod matrix. *Journal of Applied Psychology*, *67*, 500-505.

Ko, Y.-H. (1994). A search for a better Likert point-scale for mental health questionnaires. *Psychological Testing*, *41*, 55-72.

Komorita, S. S., & Graham, W. K. (1965). Number of scale points and the reliability of scales. *Educational and Psychological Measurement*, *15*, 987-995.

Likert, R. (1932). A technique for the measurement of attitudes. *Archives of Psychology*, *22*, 2-55.

Masters, J. R. (1974). The relationship between number of response categories and reliability of Likert-type questionnaires. *Journal of Educational Measurement*, *11*, 49-53.

Matell, M. S., & Jacoby, J. (1971). Is there an optimal number of alternatives for Likert scale items? Study I: Reliability and validity. *Educational and Psychological Measurement*, *31*, 657-674.

Oaster, T. R. F. (1989). Number of alternatives per choice point and stability of Likert-type scales. *Perceptual and Motor Skills*, *68*, 549-550.

Ory, J. C., & Wise, S. L. (1981, April). *Attitude change measured by scales with 4 and 5 response options*. Paper presented at the meeting of the National Council on Measurement in Education, Chicago, IL.

Preacher, K. J. (2002). Calculation for the test of the difference between two independent correlation coefficients [Computer software]. Retrieved from http://www.quantpsy.org

Preston, C. C., & Colman, A. M. (2000). Optimal number of response categories in rating scales: Reliability, validity, discriminating power, and respondent preferences. *Acta Psychologica*, *104*, 1-15.

Reid, J. (1990). The dirty laundry of ESL survey research. *TESOL Quarterly*, *24*, 323-338.

Scott, J. (1997). Children as respondents: Methods for improving data quality. In L. Lyberg, P. Biemer, M. Coolins, E. D. deLeeuw, C. Dippo, N. Schwarz, et al. (Eds.), *Survey measurement and process quality* (pp. 331-350). New York, NY: Wiley.

Symonds, P. M. (1924). On the loss of reliability in ratings due to coarseness of the scale. *Journal of Experimental Psychology*, *7*, 456-461.

Weng, L.-J. (2004). Impact of the number of response categories and anchor labels on coefficient alpha and test-retest reliability. *Educational and Psychological Measurement*, *64*, 956-972.

Wong, C.-S., Chuen, K.-C., & Fung, M.-Y. (1993). Differences between odd and even number of response scales: Some empirical evidence. *Chinese Journal of Psychology*, *35*, 75-86.

Zuo, G. Y. (2007). Toward using confidence intervals to compare correlations. *Psychological Methods*, *12*, 399-413.