SELECTING THE NUMBER OF RESPONSE CATEGORIES FOR A LIKERT-TYPE SCALE

Nicholas J. Birkett, McMaster University

ABSTRACT

On theoretical grounds, one would expect there to be a positive correlation between reliability and the number of response categories used in a Likert-type scale. Many psychometrists recommend that at least 20 response categories be used. However, when respondents are presented with either too many or too few response categories, it is possible that respondent fatigue might occur with a corresponding drop-off in response rate and reliability. This hypothesis was tested by comparing the results of three simultaneous surveys done using comparable versions of the Multidimensional Health Locus of Control Questionnaire employing two, six, and fourteen response categories. These were mailed to independent samples of 300 randomly selected adults. There was no significant correlation between the number of response categories and the response rate. Significant differences in reliability were found. Reliability tended to be highest with the questionnaire having six response categories. These data support the practice of employing about seven response categories.

INTRODUCTION

Epidemiologists often use questionnaires which incorporate rating scales as indices of a person's functioning. Examples include psychosocial function [1,2], health locus of control [3] and measures of patient satisfaction [4]. A common method of scale construction for indices uses a multiple point Likert-type scale [5, P210]. Evaluation of an instrument's properties should include consideration of the reliability with which it measures the construct in question. The reliability of a rating scale measures the extent to which raters provide consistent results on repeat measurement. Two common methods of measuring reliability are: test-retest (administer the scale twice to the same respondent and compare the ratings) and internal consistency (compare responses to similar items on the scale).

One goal of developing a new measuring instrument is to maximize the reliability of the index. It has been well established that increasing the number of items in an index will increase reliability [5, P210]. In addition to determining the number of items in the scale, one must also select the number of response categories to provide for each item. The effect of this choice on reliability is less clear. A number of publications examining this effect have been published in the psychological and marketing literature [6-16] but this design issue has not received attention in the epidemiologic literature.

One approach to determining the optimal number of response categories is based on a theoretical model wherein it is assumed that the responses have an underlying Gaussian distribution for each respondent. The true, continuous response is categorized into a smaller number of allowable responses (the Likert-scale response categories). Examination of this model using algebraic techniques [6] and Monte Carlo simulation [7-9] reveals a monotonic increase in reliability as the number of response categories increase. However, the increase in reliability becomes small after five to seven categories are used.

The conclusions of the modeling approach are dependent on the validity of the underlying model. Consequently, a number of empirical studies have also been conducted [10-16]. In general, these studies either found no relation between reliability and the number of response categories or demonstrated an inverted U-shape pattern with maximum reliability occuring with between five and seven response categories. Komorita and Graham [12] suggested that scales having heterogenous items would show a positive correlation between reliability and the number of response categories but this was not confirmed by Masters [15] or McKelvie [16]. Masters [15] found a positive correlation when responses were consistent across respondents.

All of these empirical studies have been conducted under controlled conditions on volunteer groups, usually psychology students. Hence, the results may not be generalizable to other populations. Further, in these studies, it was not possible to examine a possible secondary effect of altering the number of response categories - a drop in response rate. It could be hypothesized that too many categories would lead to respondent fatigue or confusion which might reduce the response rate and effectively counter-balance any increase in reliability.

This paper presents the results of a randomized evaluation of the effect of varying the number of response categories on the reliability and response rate to a mailed questionnaire sent to a community-based probability sample.

METHODS

Study participants were selected by random sampling from the tax assesment roles of the city of London, Ontario, Canada. The tax assessment role enumerates the name, age and sex of all individuals living in, or paying taxes to, London. Three independent random samples of 450 entries were selected from this frame using a computer generated set of random numbers. After eliminating ineligible children and businesses, a further random sub-sample was selected to produce three final study groups of 300 individuals each.

The study questionnaire was the Multidimensional Health Locus of Control Questionnaire (MHLC) of Wallston and Wallston [3]. This instrument had been selected for use in a community-based hypertension survey [17]. The MHLC contains 18 items with 3 subscales [appendix 1]. It measures the extent to which the respondent believes that their health is controlled by three sources: internal factors (e.g. exercise), powerful others (e.g. physicians) and chance. The standard version of the MHLC provides for 6 labelled response categories (strongly disagree, moderately disagree, slightly disagree, slightly agree, moderately agree, strongly agree). For our study, we developed two alternate forms. The only difference between the forms was in the number of response categories provided - Form 1 had two categories (agree, disagree); Form 2 had six categories; and Form 3 had fourteen categories (very strongly agree, stongly agree, moderately agree, between moderately and mildly agree, mildly agree, slightly agree, very slightly agree, etc.). The three study samples were randomly assigned to receive one of these three forms in a mail survey. The subjects were not informed of the methodologic evaluation being performed. Each questionnaire included a pre-paid, pre-addressed return envelope. After two weeks, a second copy of the questionnaire was sent to all non-responders. No further attempt was made to increase the response rate.

Respondent eligibility was examined further when the questionnaires were returned. Subjects were ruled ineligible if they had died or were no longer living at the target address since they should not have been included in the original sampling frame.

The MHLC questionnaires were scored as described by Wallston et.al. [3] by summing items values across subscales. Three subscale scores were obtained: internal, chance and powerful others. Reliabililty was measured using Cronbach's alpha coefficient [18]. Since Cronbach's alpha can be interpreted as an intra-class correlation coefficient [18], the method of Kraemer [19] was used to test the null hypothesis of equal reliability in the three groups. This method produces a likelihood ratio statistic which is asymptotically distributed as a Chi-square statistic with appropriate degrees of freedom. Response rates to the first mailing and to the combination of both mailings were estimated by dividing the total number of responses by the number of eligible respondents. Second mailing response rate used the number of eligible second mailings as the denominator. Response rates were compared using a Chi-square analysis of the 2x3 contingency table. Logistic regression was used to control for age and sex differences between the groups.

RESULTS

Sample Characteristics

The basic characteristics of the three samples are summarized in Table I. Overall, 21% of the sample was ineligible. This ranged from 18% to 23%. Forty-eight percent (48%) of the total sample was male with a range from 47% to 50%. The mean age was 41 years. The age and sex characteristics were similar to those found on the 1981 London census. There were no significant differences amongst the three groups on eligibility and sex. However, a one-way ANOVA revealed a marginally significant age difference (F=3.14, 2 and 712 degrees of freedom (d.f.), p=0.04). Hence, age-adjusted rates will also be considered in further analysis.

Response Rate

Overall response rate to the first mailing was 50% with an increase to 67% after the second mailing. The response rate to both mailings within each sample is shown in Table II. There were no significant differences in response rate among the three samples. Age and sex effects on response rates to the first and second mailings were examined by calculating the response rate in both sexes and in two groups using 40 years of age as a dividing criterion. Twelve tests were conducted but only one was significant at the 5% level (second mailing; age over 40; X^2=7.0, 2 df, p=0.03). A multivariate stepwise logistic regression was performed to adjust for any residual age/sex confounding. Separate analyses were performed for response rate to the first mailing, to the second mailing and to both mailings. Age, sex and sample membership were not significant in any of the regressions.

Reliabililty

The reliability of the three MHLC subscales was measured by Cronbach's alpha and is shown in Table III for each group. The null hypothesis of equal reliability was rejected for the Powerful Others Subscale (p=0.02) while for the Chance and Internal scales it was not rejected. Reliability was also examined in the same four age and sex groups as for response rate. No significant differences were found (p>0.05). However, in six of the twelve tests, group two had the highest reliability.

Each sub-scale was composed of six itmes. Hence, according to the Spearman-Brown Prophecy Formula, reliability for the entire sub-scale would be increased over that for the individual items [5,P210]. Thus, it is possible that increasing the number of response categories could increase reliability but that this effect would be masked in a six item scale. In an effort to examine this effect, each of the three sub-scales was partitioned into three, two-item sub-scales by pairing consecutive items. Reliability was recalculated for the new two-item sub-scales (Table IV). Four of the nine tests were significant (p<0.05) and two more were close to significant (p<0.10). The overall mean reliabilities among the three groups were not significantly different (X^2=0.31, 2 df, p>0.10). ANOVA techniques were also applied to the reliability data after transformation by Fisher's Z-transformation ($r'=\ln((1+p)/(1-p))$). The effect of the number of response categories was not significant (F = 1.1, 2 and 8 df, p>0.10).

DISCUSSION

Based on the results of theoretical modeling and Monte Carlo simulation, it was expected that the reliability of the test questionnaire would increase monotonically as the number of response categories increased. The results from Tables III and IV are not consistent with this hypothesis. Rather, there is a tendency for the reliability to fall off when fourteen response boxes are used. This could reflect respondent confusion/fatique and is consistent with the original hypothesis underlying this work and with published results from volunteer subjects.

Contrary to expectations, no significant effect on response rate was noted. One marginally statistically significant effect was noted but, since multiple tests were conducted, caution must be used in interpreting this result. Response rate tended to be higher in sample three (fourteen response categories) but was also higher in older persons and sample three contained more older subjects. Simultaneous adjustment of age, sex and number of response categories did not reveal any significant results.

When a statistical analysis does not reject the null hypothesis, one must consider the possibility of a Type II error and determine the power of the study against alternate hypotheses of interest. Using the standard formula for sample size when comparing proportions, we can determine that this study had a 90% power to detect a change in response rate of 15% and a 60% power to detect a 10% change in response rate. Hence, this study can rule out the possibility of a large change in response rate but some relevent changes in response rate might have been missed. Evidence against this possibility is provided by examining the trends in response rates amongst the three samples. There is no trend consistent with the initial hypothesis of a lower response rate with the fourteen response category questionnaire. In fact, the observed trend is reversed from the hypothesized effect.

This study only examined one instrument administered in a mail survey. As usual, caution should be exercised in generalizing our results to other settings. The MHLC is a short instrument (18 items). It is possible that different results might be found if a longer or more complicated instrument were employed. However, the consistency of our results with those on volunteer subjects using different instruments is encouraging.

Present epidemiological practice usually leads to the construction of scales with five or seven categories. The results of this study are consistant with this practice. Theoretical studies suggest that the further increase in reliability by providing more than seven response categories would be slight. Therefore, the present practice is a reasonable compromise between subject acceptability and theoretical optimality.

Acknowledgements

This work was conducted while Dr. Birkett held a Research Scientist Award from the Ontario Government and was funded by a research allowance associated with this award. The author thanks Mrs. J. Schieven for conducting the data collection and Mrs. M. Taylor for typing this manuscript. Dr's.Streiner, Guyatt and Norman provided helpful comments to an earlier draft of this manuscript.

BIBLIOGRAPHY

1. Cottington EM, Brock BM, House JS, Hawthorne VM. Psychosocial factors and blood pressure in the Michigan statewide blood pressure survey. Amer J Epid 1985;121:515-529.
2. Joffres M, Reed DM, Nomura AMY. Psychosocial processes and cancer incidence among Japanese men in Hawaii. Amer J Epid 1985;121:488-500.
3. Wallston KA, Wallston BS. Development of the multidimensional health locus of control (MHLC) scales. Health Education Monographs 1978;6:160-170.
4. Zyzanski SJ, Hulka BS, Cassel JC. Scale for the measurement of "satisfaction" with medical care: modifications in content, format and scoring. Med Care 1974;7:611-620.
5. Nunally JC. Psychometric Theory, Second Edition. New York, NY, McGraw-Hill Book Company, 1978.
6. Guilford JP. Psychometric Methods. New York, N.Y., McGraw-Hill Book Company, 1954.
7. Green PE, Rao VR. Rating scales and information recovery - how many scales and response categories to use? J Marketing 1970;34:33-39.
8. Nishisato S, Torii Y. Effects of categorizing continuous normal variables on product-moment correlation. Jap Psychol Res 1970;13:45-49.
9. Lissitz RW, Green SB. Effect of the number of scale points on reliability: a Monte Carlo approach. J Appl Psychol 1975;60:10-13.
10. Bendig AW. The reliability of self-ratings as a function of the amount of verbal anchoring and of the number of categories on the scale. J Appl Psychol 1953;37:38-41.
11. Bendig AW. Reliability and the number of rating scale categories. J Appl Psychol 1954;38:38-40.
12. Komorita SS, Graham WK. Number of scale points and the reliability of scales. Ed Psychol Measur 1965;25:987-995.
13. Matell MS, Jacoby J. Is there an optimal number of alternatives for Likert scale items? Study I: Reliability and Validity. Ed Psychol Measur 1971;31:657-674.
14. Finn RH. Effects of some variations in rating scale characteristics on the mean and reliabilities of ratings. Ed Psychol Measur 1972;32:255-265.
15. Masters JR. The relationship between number of response categories and reliability of Likert-type questionnaires. J Ed Measur 1974;11:49-53.
16. McKelvie SJ. Graphic rating scales - how many categories? Br J Psychol 1978;69:185-202.
17. Birkett NJ, Donner A, Maynard M. Health locus of control in a cross-sectional hypertension survey. Presented at: National Conference on High Blood Pressure Control, 1985.
18. Cronbach LJ. Coefficient alpha and the internal structure of tests. Psychometrika 1951;16:297-334.

19. Kraemer HC. On estimation and hypothesis testing problems for correlation coefficients. Psychometrika 1975;40:473-485.

APPENDIX 1

MULTIDIMENSIONAL HEALTH LOCUS OF CONTROL SCALE

Internal Subscale

1. If I get sick, it is my own behavior which determines how soon I get well again.
6. I am in control of my health.
8. When I get sick I am to blame.
12. The main thing which affects my health is what I myself do.
13. If I take care of myself, I can avoid illness.
17. If I take the right actions, I can stay healthy.

Chance Subscale

2. No matter what I do, if I am going to get sick, I will get sick.
4. Most things that affect my health happen to me by accident.
9. Luck plays a big part in determining how soon I will recover from an illness.
11. My good health is largely a matter of good fortune.
15. No matter what I do, I'm likely to get sick.
16. If it's meant to be, I will stay healthy.

Powerful Others Subscale

3. Having regular contact with my physician is the best way for me to avoid illness.
5. Whenever I don't feel well, I should consult a medically trained professional.
7. My family has a lot to do with my becoming sick or staying healthy.
10. Health professionals control my health.
14. When I recover from an illness, it's usually because other people (for example, doctors, nurses, family, friends) have been taking good care of me.
18. Regarding my health, I can only do what my doctor tells me.

TABLE 1

SAMPLE CHARACTERISTICS

		2 Response Boxes	6 Response Boxes	14 Response Boxes	Total
Ineligible		71 (23%)	66 (22%)	55 (18%)	192 (21%)
Eligible		233 (77%)	236 (78%)	246 (82%)	715 (79%)
GENDER:	Male	142 (47%)	151 (50%)	143 (47%)	436 (48%)
	Female	158 (52%)	149 (49%)	157 (52%)	464 (51%)
	Unknown	4 (1%)	2 (1%)	1 (1%)	7 (1%)
AGE:	Mean (yrs)	38.7	41.7	42.2	40.9
	s.d.	16.4	18.8	17.4	17.6

Eligibility: $X^2 = 2.57$, 1 d.f., $p = 0.29$
Gender: $X^2 = 2.63$, 2 d.f., $p = 0.27$
Age: $F = 3.14$, 2 & 712 d.f., $p = 0.04$

TABLE 2

EFFECT OF NUMBER OF RESPONSE CATEGORIES ON RESPONSE RATE

MAILING	2 Response Boxes	6 Response Boxes	14 Response Boxes	Combined	Signif.*
FIRST	119 (51%)	118 (50%)	122 (50%)	359 (50%)	$X^2 = 0.11$
SECOND	36 (32%)	34 (29%)	51 (41%)	121 (34%)	$X^2 = 4.52$
BOTH	155 (67%)	152 (64%)	173 (70%)	480 (67%)	$X^2 = 1.97$
Total sample size	233	236	246	715	

* Comparison of three samples, using a Chi-square test with 2 degrees of freedom.

TABLE 3

EFFECT OF NUMBER OF RESPONSE CATEGORIES ON THE RELIABILITY OF THE SUB-SCALES

Sub-scale	2 Response Boxes	6 Response Boxes	14 Response Boxes	Significance+
Internal	.680	.681	.725	$X^2 = 0.79$
Chance	.613	.651	.616	$X^2 = 0.35$
Powerful Others	.644	.731	.543	$X^2 = 7.72$ *
Sample size	149	145	168	

+ Comparison of three samples using Kraemer's test for intra-class correlation coefficients which yields a Chi-square test with 2 degrees of freedom.

* $0.01 < p < 0.05$

TABLE 4

EFFECT OF NUMBER OF RESPONSE CATEGORIES ON THE RELIABILITY OF TWO ITEM SUB-SCALES

Sub-scale	2 Response Boxes	6 Response Boxes	14 Response Boxes	Significance+
Int (Q1,6)	.366	.391	.568	$X^2 = 6.53$ *
Int (Q8,12)	.221	.413	.441	$X^2 = 5.55$
Int (Q13,17)	.741	.640	.638	$X^2 = 3.88$
Chance (Q2,4)	.241	.136	-.043	$X^2 = 6.76$ *
Chance (Q9,11)	.499	.464	.459	$X^2 = 0.25$
Chance (Q15,16)	.500	.375	.416	$X^2 = 1.86$
Pow Oth (Q3,5)	.570	.693	.468	$X^2 = 9.32$ **
Pow Oth (Q7,10)	-.072	.180	.018	$X^2 = 4.86$
Pow Oth (Q14,18)	.614	.513	.376	$X^2 = 8.09$ *
Mean	.409	.423	.371	$X^2 = 0.31$
Sample size	149	145	168	

+ Comparison of three samples using Kraemer's test for intra-class correlation coefficients which yields a Chi-square test with 2 degrees of freedom.

* $0.01 < p < 0.05$
** $0.001 < p < 0.01$