ELSEVIER

Acta Psychologica 104 (2000) 1–15

acta psychologica

www.elsevier.com/locate/actpsy

Optimal number of response categories in rating scales: reliability, validity, discriminating power, and respondent preferences

Carolyn C. Preston [a], Andrew M. Colman [b,*]

[a] *Department of General Practice and Primary Health Care, University of Leicester, University Road, Leicester LE1 7RH, UK*

[b] *Department of Psychology, University of Leicester, University Road, Leicester LE1 7RH, UK*

Received 13 April 1999; received in revised form 13 September 1999; accepted 14 September 1999

Abstract

Using a self-administered questionnaire, 149 respondents rated service elements associated with a recently visited store or restaurant on scales that differed only in the number of response categories (ranging from 2 to 11) and on a 101-point scale presented in a different format. On several indices of reliability, validity, and discriminating power, the two-point, three-point, and four-point scales performed relatively poorly, and indices were significantly higher for scales with more response categories, up to about 7. Internal consistency did not differ significantly between scales, but test–retest reliability tended to decrease for scales with more than 10 response categories. Respondent preferences were highest for the 10-point scale, closely followed by the seven-point and nine-point scales. Implications for research and practice are discussed.

PsycINFO classification: 2220

Keywords: Item analysis; Questionnaires; Rating scales; Test reliability; Test validity

* Corresponding author. Tel.: +44-1-16-2522167; fax: +44-1-16-2522067.
E-mail address: amc@le.ac.uk (A.M. Colman).

0001-6918/00/$ - see front matter
PII: S0001-6918(99)00050-5

1. Introduction

Rating scales are among the most widely used measuring instruments in psychology, and it is therefore not surprising that a great deal of research has been devoted to the effects of variations in rating scale format, including differences in the number of response categories. In current practice, most rating scales, including Likert-type scales and other attitude and opinion measures, contain either five or seven response categories (Bearden, Netmeyer, & Mobley, 1993; Peter, 1979; Shaw & Wright, 1967).

In spite of decades of research, the issue of the optimal number of response categories in rating scales is still unresolved. Some investigators have studied response patterns and information retrieval. Schutz and Rucker (1975) found in their study of response patterns that "the number of available response categories does not materially affect the cognitive structure derived from the results" (p. 323), which seems to suggest that the number of response categories has little effect on the results obtained. This conclusion is not in line with the findings of other studies, which have provided support for the use of scales with more than two or three response categories. For example, Garner (1960) suggested that maximum information is obtained by using more than 20 response categories. Green and Rao (1970), on the other hand, found that information retrieval is maximized by using six or seven response categories, with little extra information being gained by increasing the number of categories beyond seven.

Symonds (1924) was the first to suggest that reliability (in this case inter-rater reliability) of scores is optimized by the use of seven categories. This suggestion was contested by Champney and Marshall (1939), who advocated the use of finer scales, but the case for seven-point scales was strengthened by Miller (1956), who suggested in an influential article that the human mind has a span of apprehension capable of distinguishing about seven different items (plus or minus two), which implies a limit of about seven on the number of categories that people are able to use in making judgments about the magnitudes of unidimensional stimuli. This has implications for rating scales: the limit on the human span of apprehension suggests that little if any additional information can be obtained by increasing the number of response categories beyond about seven. The reliability of scores derived from scales with different numbers of response categories was later investigated by Bendig (1953, 1954), who found relatively constant test–retest reliabilities over scales with two, three, five, seven, and nine response categories, relatively constant inter-rater reliability over scales with three, five, seven, and nine response categories, and a decrease in reliability for 11-point scales. A few subsequent researchers confirmed Bendig's finding that reliability is largely independent of the number of response categories (e.g., Boote, 1981; Brown, Wilding, & Coulter, 1991; Komorita, 1963; Matell & Jacoby, 1971; Peabody, 1962; Remington, Tyrer, Newson-Smith, & Cicchetti, 1979).

Some researchers in this area have arrived at different conclusions regarding reliability. In a study based on Monte-Carlo simulation methods, Cicchetti, Showalter and Tyrer (1985) found evidence for an increase in inter-rater reliability from two-point to seven-point scales; beyond this – even up to 100 response categories – no

substantial increase in reliability was found. These researchers concluded that "the differences in scale reliability between a 7-, 8-, 9-, or 10-category ordinal scale on one hand, and a 100-point or continuous scale on the other is trivial . . . 7 ordinal categories of response appear at least functionally interchangeable with as many as 100 such ordered categories" (p. 35). Similar conclusions were drawn by Oaster (1989) with regard to test–retest reliability and inter-item consistency, and a number of other researchers have reported that reliability is maximized with seven-point scales (Finn, 1972; Nunnally, 1967; Ramsay, 1973). These studies provide support for the early findings of Symonds (1924), mentioned above, and more generally for the continued popularity of seven-point scales in practice (see Cox, 1980, for a review). A few researchers have, however, reported higher reliabilities for five-point scales (Jenkins & Taber, 1977; Lissitz & Green, 1975; McKelvie, 1978; Remmers & Ewart, 1941), and a recent study using the multitrait-multimethod approach found evidence for higher monotrait-monomethod (MTMM) reliability in four-point than six-point scales (Chang, 1994).

In a comparatively small number of studies, validity has been used as a criterion for judging the performance of scales with different numbers of response categories. Matell and Jacoby (1971) carried out a thorough empirical study comparing scales with varying numbers of response categories (from 2 to 19) and concluded that as few as two response categories may be adequate in practice. They suggested that both reliability and validity are independent of the number of response categories, and their results implied that collapsing data from longer scales into two-point or three-point scales would not diminish the reliability or validity of the resulting scores. Loken, Pirie, Virnig, Hinkle and Salmon (1987) examined the criterion validity of various scales through their ability to differentiate between different population groups and found 11-point scales to be superior to three-point or four-point scales. Hancock and Klockars (1991) found that nine-point scale scores correlated better than five-point scale scores with objective measures of the original stimuli. In a comparison using a MTMM covariance matrix, Chang (1994) found approximately similar criterion validity coefficients for four-point and six-point scales but higher convergent validity coefficients for the six-point scales. Taken together, these studies tend to suggest that validity increases with increasing numbers of response categories or scale points.

Respondent preferences have not been investigated in depth in previous studies of rating scales. However, Jones (1968) examined respondents' preferences for scales with two or seven response categories and a graphic rating scale and reported that the dichotomous scale was judged to be less "accurate", less "reliable", less "interesting", and more "ambiguous" than both the seven-point and the graphic rating scales, but the two-point and graphic rating scales were judged to be easier to use. Respondents expressed a clear preference for multiple-category over dichotomous scales.

The aim of the investigation reported below is to provide a thorough assessment, using multiple independent criteria, of the reliability, validity, and discriminating power of scores from rating scales varying widely in number of response categories. A secondary aim is to investigate the other important issue in rating scale design,

namely respondent preferences. Our research is more detailed and thorough than most previous studies in the area, and its design allows us to examine not only several objective indices of reliability, validity, and discriminating power, but also subjective measures of respondents' opinions about the scales. For example, if a scale is too difficult to use, or too simple to allow respondents to express themselves, then respondents may become frustrated and demotivated, and the quality of their responses may decrease. In addition to considerations of reliability, validity, and discriminating power, a test designer may wish to take respondent preferences into account when constructing a rating scale.

1.1. Materials and methods

A questionnaire was administered to 149 respondents (45 males and 104 females, aged from 18 to over 60 with a mode of 20), the majority of whom (134) were undergraduate students at the University of Leicester. The respondents were recruited by a form of snowball sampling in which students who volunteered to participate as respondents also recruited additional respondents in return for course credits. The sample thus consisted of undergraduate students and their friends (some of whom were also undergraduate students) and relatives.

Each respondent filled in a questionnaire consisting of a number of rating scales with varying numbers of response categories relating to the quality of service provided by either a store or a restaurant of which he or she had personal experience. These two service categories were chosen on the assumption that all respondents would have visited a store or restaurant in the recent past, and this turned out to be the case. Respondents were first required to provide a global rating of their chosen store or restaurant for overall quality from 0 (*extremely bad*) to 100 (*extremely good*); this global rating served as a criterion measure for later assessments of validity. The rest of the questionnaire consisted of 11 sets of five questions each, in which respondents rated the quality of five service elements for their chosen store or restaurant. To avoid order-of-presentation effects that might have arisen from using a single random order of presentation, the 11 sets were presented in a different randomized order to each respondent. For the restaurant, the five service elements were "competence of staff", "promptness of service", "range of choice", "cleanliness of premises", and "individual attention"; for the store, the service elements were "helpfulness of staff", "promptness of service", "range of products", "tidiness of premises", and "individual attention".

The 11 sets of questions to which the participants responded were identical apart from the number of response categories in the rating scales associated with each of the five service elements in each set. The number of response categories of the rating scales were 2, 3, 4, 5, 6, 7, 8, 9, 10, and 11. Each scale was end-anchored with *very poor* on the left and *very good* on the right and was presented as a series of numbers spaced out across the page representing the response categories. Respondents were required to circle the appropriate number in each case depending on their opinions or judgments. A 101-point scale was also included, with respondents being asked to rate each service element from 0 (*very poor*) to 100 (*very good*), but in this case no

numbers were displayed: respondents were required simply to write down appropriate numbers between 0 and 100 reflecting their ratings.

In order to measure respondents' preferences for the various scales, they were also asked to rate each scale on three different aspects of scale performance. Using the 101-point scale described above, from 0 (*very poor*) to 100 (*very good*), they rated each scale on the following three criteria: "ease of use", "quick to use", and "allowed you to express your feelings adequately".

In order to provide data for an assessment of test–retest reliability, each respondent completed an identical retest questionnaire between one and three weeks after the original testing – the second questionnaire was issued to the respondents one week after the first and had to be returned within two weeks from that date. Of the 149 respondents who completed the first questionnaire, 129 completed the second within the two-week time limit, and the retest results are based on their responses. This response rate of 86% is high for research of this type.

2. Results

2.1. Preliminary checks

The psychometric criteria that are used to compare scales with different numbers of response categories make sense only if such scales measure the same construct. We therefore carried out three preliminary checks.

First, we used a procedure described by Green (1992) to determine whether the 12 correlation matrices representing the associations between the five items in each scale, one correlation matrix per scale length, differed significantly from one another. The weighted least squares chi-square value was 144.09, $df = 155, p = 0.72$, suggesting no evidence of significant differences between the correlation matrices.

Second, the unnormed and normed fit indices (Bentler & Bonnet, 1980) that were achieved were 1.00 and 0.96, respectively; according to Bentler and Bonnet (1980), this indicates that little additional increase in goodness of fit would be likely to be achieved by allowing some of the correlation matrices to differ from others.

Third, we performed a maximum-likelihood factor analysis on each of the 11 correlation matrices and then tested whether the ordinal structure of the factor loadings was the same across matrices. The rank-order of the factor loadings within each scale length (matrix) turned out to be identical across all 11 matrices, yielding a Kendall coefficient of concordance of 1.00.

All three of these preliminary checks confirm that the scales with different numbers of response categories all measure the same underlying construct.

2.2. Reliability

Ratings derived from each scale were evaluated for test–retest reliability (stability) and also, using Cronbach's alpha, for consistency over the five questions within each scale type (internal consistency reliability). Table 1 shows the reliability coefficients

Table 1
Reliability of rating scales with different numbers of response categories

Test–retest	Response categories										
	2	3	4	5	6	7	8	9	10	11	101
Reliability	0.88	0.86	0.89	0.91	0.92	0.93	0.94	0.94	0.93	0.92	0.90
Cronbach's α	0.81	0.79	0.82	0.82	0.83	0.85	0.85	0.85	0.85	0.86	0.85

for the test–retest reliability analysis and the alpha coefficients for the internal consistency reliability analysis.

The reliability coefficients are all relatively high (0.79 or above) and statistically significant beyond $p < 0.05$, and the effect sizes are all large according to Cohen's (1988, 1992) criterion. Test–retest reliability coefficients were lowest for two-point, three-point, and four-point scales and highest for scales with about 7 to 10 response categories; there was a slight decline in test–retest reliability coefficients for scales with 11 and 101 response categories. Statistical tests for the significance of differences between correlations were carried out using Fisher's r to z transformation (Howell, 1992, p. 251). Statistically significant at $p < 0.05$ were the differences between the two-point scale and the scales with 6, 7, 8, 9, and 10 response categories; between the three-point scale and the scales with 6, 7, 8, 9, 10, and 11 response categories; between the four-point scale and the eight-point and nine-point scales; and between the 101-point scale and the eight-point and nine-point scales. All other differences between the test–retest reliability coefficients were statistically nonsignificant.

Cronbach alpha coefficients were lowest for two-point and three-point scales, and like the test–retest reliability coefficients they increased with increasing numbers of response categories up to seven. Little further increase in reliability was found above seven response categories: alpha coefficients for scales with 7, 8, 9, 10, 11, and 101 response categories were all very similar. Using Feldt's test for the significance of differences between more than two related alpha coefficients (Woodruff & Feldt, 1986), it was determined that none of the differences between the alpha coefficients were statistically significant: using either of the recommended statistics HAN2 or UX1, $\chi^2(10) < 18.31$, $p > 0.05$.

2.3. Validity and discriminating power

Validity and discriminating power were assessed in several different ways. First, an index of the criterion validity of scores derived from each scale was created by computing an aggregate score for each respondent over the questions relating to the five service elements and then correlating these aggregate scores with the respondents' scores on the criterion global rating of overall quality of service. The global rating of overall service quality was included in the same questionnaire as the ratings of particular service elements, but it was separated from these other more specific measures. There are many precedents for the use as validity criteria of scales that were included in the same questionnaire as the scales in question (Althauser,

Heberlein, & Scott, 1971; Campbell & Fiske, 1959; Messick, 1993). The global rating that was used in this study was chosen because there was no genuinely *external* criterion available, and because it is reasonable to assume that ratings of particular service elements, if they measure what they purport to measure, ought to correlate with global ratings of overall service quality.

Next, an index of the intertertile discriminating power of scores from each scale was obtained by calculating a t statistic relating to the mean difference in aggregate scores for the scale's five questions between the tertile of respondents who provided the highest global ratings of overall service quality (over 80 on the scale from 0 to 100) and the tertile who provided the lowest global quality rating (0–25). To provide a second index of the discriminating power of each scale, item-whole correlations were calculated between the ratings of each of the five service elements and the aggregate score from all scales relating to the corresponding service element. For each scale, the index of discriminating power was provided by the mean item-whole correlation over the five questions in the scale. For these calculations, ratings from all scales apart from the 101-point scale were rescaled to make them comparable using the following formula:

$$(\text{rating} - 1)/(\text{number of response categories} - 1) \times 100.$$

Finally, convergent validity (Campbell & Fiske, 1959) was evaluated by examining the correlations of scores on each scale with scores on each of the others. Scores from a scale were assumed to show convergent validity to the extent to which they correlated with scores from other scales measuring the same underlying construct (see also Althauser et al., 1971; American Psychological Association, 1985, pp. 9–10; Chang, 1994; Messick, 1993).

Data concerning criterion validity, intertertile discriminating power, and item-whole correlations are presented in Table 2, and data concerning convergent validity are presented in Table 3.

The results in Table 2 show that the pattern for validity and discriminating power is similar to the pattern that was found for reliability. Correlations between scale scores and the criterion variable, t values for intertertile discriminating power,

Table 2
Criterion validity, intertertile discriminating power, and item-whole correlations[a]

	Response categories										
	2	3	4	5	6	7	8	9	10	11	101
Criterion											
Validity (r)	0.83	0.82	0.85	0.87	0.88	0.87	0.87	0.89	0.87	0.88	0.89
Intertertile											
Discrim. (t)	19.2	16.8	18.6	20.8	21.5	20.8	22.2	23.7	22.3	23.0	23.4
Item-whole (r)	0.87	0.90	0.92	0.95	0.96	0.97	0.97	0.96	0.96	0.96	0.96

[a] All t statistics are significant at $p < 0.0001$.

Table 3
Convergent validity: intercorrelations between scales with different numbers of response categories

Categories	2	3	4	5	6	7	8	9	10	11	101
2											
3	0.836										
4	0.859	0.878									
5	0.857	0.906	0.921								
6	0.884	0.899	0.928	0.956							
7	0.884	0.907	0.922	0.956	0.969						
8	0.893	0.908	0.923	0.949	0.973	0.964					
9	0.882	0.907	0.913	0.959	0.965	0.969	0.972				
10	0.880	0.903	0.905	0.955	0.959	0.964	0.975	0.977			
11	0.871	0.902	0.905	0.950	0.958	0.963	0.966	0.970	0.974		
101	0.883	0.910	0.902	0.933	0.954	0.956	0.962	0.964	0.952	0.959	

and item-whole correlations were statistically significant for all scales, and the effect sizes are all large according to Cohen's (1988, 1992) criteria. As regards criterion validity, scores from the scales with two, three, and four response categories produced the lowest correlations (below $r = 0.86$) with the criterion. Correlations for scores from 5-point to 101-point scales were higher and very similar to each other (around $r = 0.88$). The scales that yielded scores with the highest criterion validity coefficients ($r = 0.89$) were those with 9 and 101 response categories. However, none of the criterion validity coefficients shown in Table 2 differs significantly at $p < 0.05$ from any of the others according to Williams's t statistic (Howell, 1992, p. 254).

Turning to intertertile discriminating power, scales with two, three, and four response categories again performed least well, with scores from the three-point scale showing the lowest intertertile discriminating power of all $g < 20$). Scores from the other scales showed similar discriminating power to one another, with nine-point, 11-point, and 101-point scales performing best. The significance of the differences between the intertertile discriminating power t values was tested by transforming the t values to z scores (Howell, 1992, p. 251). The only statistically significant differences at $p < 0.05$ were between the t values for the three-point scale on the one hand and the scales with nine and 101 response categories on the other.

Item-whole correlations were also lowest for the scales with two, three, or four response categories and increased with increasing numbers of response categories up to about six. There was no substantial change in item-whole correlation coefficients for the scales with more than six response categories. Fisher's r to z transformation was used to evaluate the significance of differences between the item-whole correlations. The coefficient for the two-point scale differed at $p < 0.05$ from all of the others. Coefficients for the three-point and four-point scales also differed at $p < 0.05$ from all the others but did not differ statistically significantly from each other. All other differences were nonsignificant.

Table 3 shows the intercorrelations between scores from scales with different numbers of response categories. Every scale correlated highly and statistically significantly with each of the others, and all of the correlation coefficients represent large effect sizes according to Cohen's (1988, 1992) criterion. These results, which provide evidence of convergent validity, also indicate that the scales with relatively more response categories (six or more) correlated best with one another, and that the two-point and three-point scales correlated less highly with the longer scales.

2.4. Respondent preferences

Using a 101-point scale described above, respondents rated each scale for its "ease of use", whether it was "quick to use", and whether it "allowed you to express your feelings adequately". Analysis of variance was carried out on the ratings given in response to each of these questions. The mean score for each scale on each of the three questions designed to measure respondent preferences, and the associated standard deviations, are shown in Table 4, along with the corresponding F values. The effect sizes were all large according to Cohen's (1988, 1992) criterion: "ease of

Table 4
Respondents' preferences for scales: mean ratings (0 = very poor, 100 = very good), and standard deviations

	Response categories											
	2	3	4	5	6	7	8	9	10	11	101	*F*
Ease of use	78.6	81.4	82.0	83.7	81.3	82.3	81.5	81.0	83.2	76.7	74.1	7.57*
(*SD*)	27.3	18.9	17.9	15.6	16.3	15.7	16.1	17.4	16.2	19.4	21.6	
Quick to use	86.6	86.8	85.5	85.1	84.5	83.5	83.1	82.1	82.9	77.8	70.6	25.23*
(*SD*)	18.9	15.8	15.2	14.2	15.0	15.5	15.9	16.6	17.0	19.6	20.5	
Express feelings	17.8	40.0	52.0	63.7	63.4	69.0	68.8	72.9	76.0	73.1	79.3	219.36*
(*SD*)	20.0	22.6	23.1	20.7	21.4	21.3	20.1	20.4	21.1	21.8	22.5	

* $p < 0.0001$.

use'', $\eta^2 = 34$; ''quick to use'', $\eta^2 = 0.63$; ''allowed you to express your feelings adequately'', $\eta^2 = 0.94$.

Table 5 shows the scales ordered from the one that received the lowest mean score (that is, the one that was rated least favorably) to the one that received the highest mean score (that is, the one that was rated most favorably) on each of the three questions related to respondent preferences. Scales in the same row that share the same subscript are *not* significantly different; all other differences between scales in the same row are statistically significant at $p < 0.05$ according to the Tukey-HSD multiple comparisons.

These respondent preference scores show several statistically significant differences between the scales. For ''ease of use'', the scales with five, seven, and 10 response categories were the most preferred, and the scales with 11 and 101 response categories were rated as least easy to use. The scales that were rated as most ''quick to use'' were the ones with the fewest response categories: the two-point, three-point, and four-point scales were rated most favorably on this criterion, and once

Table 5
Scales ranked in order of increasing respondent preference: statistically significant differences[a]

	Scales (No. of response categories)									
	Lowest									Highest
Ease	101	11	2	9_a	6_{ab}	3_b	8_b	4	7	10	5
Quick	101	11	9	10	8_a	7_a	6	5_b	4_b	2_c	3_c
Express feelings	2	3	4	6_a	5_a	8	7	9_b	11_b	10	101

[a] *Note.* Scales in the same row that share the same subscript do not differ significantly; all other differences between scales in the same row are statistically significant at $p < 0.05$ according to the Tukey-HSD comparison.

again the scales with 11 and 101 response categories were least preferred. Ratings of the degree to which the scales "allowed you to express your feelings adequately" showed the greatest differentiation between the scales. The two-point and three-point scales received extremely low ratings on this criterion (well below 50 on the 0–100 scale). In general, longer scales tended to receive more favorable ratings on this dimension: the scales with 9, 10, 11 and 101 response categories were rated highest. Taking into account all three questions related to respondent preferences, the shortest scales (two, three and four response categories) generally received the lowest ratings, but the longest scales (11-point and 101-point scales) also received relatively low ratings. The scale that scores best overall according to respondent preferences was the 10-point scale, closely followed by the seven-point and nine-point scales.

3. Discussion

The rating scales that yielded the least reliable scores turned out to be those with the fewest response categories. Test–retest reliability (stability) was lowest for two-point, three-point, and four-point scales and was significantly higher for scales with more response categories; the most reliable scores were derived from scales with 7, 8, 9, or 10 response categories. Internal consistency was lowest for scales with two or three response categories and highest for those with seven or more, although on this criterion of reliability the differences between the reliability coefficients were not statistically significant. Our results also provide evidence of a decrease in test–retest reliability for scales with more than 10 response categories, although only the decrease from the eight-point and nine-point scales to the 101-point scale attained statistical significance. Bendig (1954) found a similar decrease in inter-rater reliability, but Cicchetti et al. (1985), using a Monte-Carlo simulation rather than a fully empirical research methodology, found no decrease in reliability for long scales.

According to the indices of validity and discriminating power that we examined in this study, the scales with relatively few response categories performed worst. The criterion validity coefficients were lowest for the scales with two, three, or four response categories and were generally higher – though these differences were not statistically significant – for scales with five or more response categories. Discriminating power was lowest for the scales with two, three, or four response categories and statistically significantly higher for scales with 9 or 101 response categories. Item-whole correlations told a similar story: scales with two, three, or four response categories performed worst, and scales with six or more response categories performed generally better – the coefficient for the two-point scale was significantly lower than the coefficients for all other scales in the investigation. Finally, the table of intercorrelations between scores on all of the scales suggested that the scales with two or three response categories yielded scores with lower overall convergent validity than the others. However, it should be borne in mind that a scale with relatively few response categories tends to generate scores with comparatively little

variance, limiting the magnitude of correlations with other scales (e.g., Chang, 1994; Martin, 1973, 1978; Nunnally, 1970). This restriction-of-range effect tends to depress the convergent validity of scores from scales with few response categories, but it is worth remembering that this arises ultimately from the inherent bluntness of such scales, which also limits their usefulness for many practical psychometric purposes.

These findings provide no corroboration for Matell and Jacoby's (1971) suggestion that reliability and validity of scores are independent of the number of response categories and that nothing is gained by using scales with more than two or three response categories; but neither do they corroborate Garner's (1960) suggestion that scales with 20 or more response categories are necessarily best. As regards reliability, the results reported above tend to confirm the findings of Symonds (1924), Nunnally (1967), Green and Rao (1970), Finn (1972), Ramsay (1973), Cicchetti et al. (1985), and Oaster (1989), which suggested that reliability of scores tends to increase from two-point to six-point or seven-point scales. As regards validity and discriminating power, our results provide a remarkably consistent picture across four independent indices, tending to confirm evidence from a small number of studies using much more restricted criteria of validity (e.g., Chang, 1994; Hancock & Klockars, 1991; Loken et al., 1987) that, statistically, scales with small numbers of response categories yield scores that are generally less valid and less discriminating than those with six or more response categories.

Respondent preference ratings differed substantially between the scales, and the differences were statistically significant. Scales with 5, 7, and 10 response categories were rated as relatively easy to use. Shorter scales with two, three, or four response categories were rated as relatively quick to use, but they were rated extremely unfavorably on the extent to which they allowed the respondents to express their feelings adequately; according to this criterion, scales with 10, 11 and 101 response categories were much preferred. On the whole, taking all three respondent preference ratings into account, scales with two, three, or four response categories were least preferred, and scales with 10, 9, and 7 were most preferred.

From the multiple indices of reliability, validity, discriminating power, and respondent preferences used in this study, a remarkably consistent set of conclusions emerges. Our results provide no support for the suggestion of Schutz and Rucker (1975) that the number of response categories is largely immaterial. On the contrary, scales with two, three, or four response categories yielded scores that were clearly and unambiguously the least reliable, valid, and discriminating. The most reliable scores were those from scales with between 7 and 10 response categories, the most valid and discriminating were from those with six or more response categories or-in the case of intertertile discriminating power – those with nine or more. The results regarding respondent preferences showed that scales with two, three, or four response categories once again generally performed worst and those with 10, 9, or 7 performed best. The superiority of scales with around seven response categories is in line with Miller's (1956) theoretical analysis of human information-processing capacity and short-term memory, subsequently refined by Simon (1974) in his characterization of information "chunks". For several decades, the vast majority of

rating scales and related psychometric instruments have used five or seven response categories (Bearden et al., 1993; Peter, 1979; Shaw & Wright, 1967). In the light of our findings, there is some support for seven-point scales, but the popularity of five-point scales seems to be less justified.

Taken together, the results reported above suggest that rating scales with 7, 9, or 10 response categories are generally to be preferred. In this study, participants responded to the scales consecutively, and this may have led to factors other than scale format affecting their ratings. Respondents may, for example, have been inconsistent through carelessness or boredom arising from having to respond to similar questions again and again. However, the satisfactory levels of scale reliability, ranging from 0.79 to 0.94 (see Table 1), and the relatively high correlations between scores from different scales, ranging from 0.84 to 0.98 (see Table 3), suggest that the respondents rated carefully and consistently across scales. Also, the fact that the scales were presented in a different random order to each respondent ensures that any inconsistencies in responding could not contribute to significant differences found between scales. Caution should, however, be exercised in generalizing the conclusions beyond the types of respondents and scales used in this study.

The research reported here examined responses to real-life experiences in recent visits to stores or restaurants. Some previous studies have focused on ratings of real-life experiences (e.g., Neuman & Neuman, 1981), whereas others have involved ratings of hypothetical experiences or abstract concepts (e.g., Matell & Jacoby, 1971; Oaster, 1989) or have used computer simulations (e.g., Cicchetti, Showalter, & Tyrer, 1985). These methodological differences probably explain, in part at least, the discrepant findings reported in these different studies. However, our aim was to investigate the effects of scale length on participant ratings using real-life experiences in order to produce findings that would have practical relevance to ratings in everyday situations such as those studied in market research.

A careful reading of our results suggests that different scales may be best suited to different purposes. Circumstances may, for example, require respondents to use a rating scale under conditions of time pressure, and in such cases it may be necessary, in order to prevent the respondents from becoming frustrated and demotivated, to use five-point or even three-point scales, because our findings show that these scales are likely to be perceived by the respondents as relatively quick and easy to use. On the other hand, where considerations of face validity are regarded as paramount, it may be important for the respondents to perceive the scales as allowing them to express their feelings adequately, and in such cases 10-point scales may be most appropriate. Before deciding on the optimal number of response categories for a rating scale, researchers and practitioners may therefore need to perform a trade-off, in the light of the prevailing circumstances, between reliability, validity, discriminating power, and respondent preferences.

The findings reported in this article relate to ratings of service quality in restaurants and stores; further research in a similar vein, using objective ratings of behavior by others, self-ratings of behavior, ratings of personality traits, ratings of the quality of products, and so on, would be needed to determine the extent to which the conclusions generalize to other domains.

Acknowledgements

Preparation of this article was supported by research awards K38 and M71 from BEM Research. We are grateful to Gareth G. Jones for drawing our attention to the problem addressed in this article and to Barry Denholm, Caroline Gaynor, and Laura Gracie for help with the collection of data. We wish to thank David Stretch and Jeremy Miles for help with the data analysis.

References

Althauser, R. P., Heberlein, T. A., & Scott, R. A. (1971). A causal assessment of validity: the augmented multitrait-multimethod matrix. In H. M. Blalock (Ed.), *Causal models in the social sciences* (pp. 374–399). Chicago, IL: Aldine.

American Psychological Association. (1985). *Standards for educational and psychological testing*. Washington, DC: Author.

Bearden, W. O., Netmeyer, R. G., & Mobley, M. F. (1993). *Handbook of marketing scales: multi-item measures for marketing and consumer behavior research*. Newbury Park, CA: Sage.

Bendig, A. W. (1953). The reliability of self-ratings as a function of the amount of verbal anchoring and the number of categories on the scale. *The Journal of Applied Psychology*, *37*, 38–41.

Bendig, A. W. (1954). Reliability and the number of rating scale categories. *The Journal of Applied Psychology*, *38*, 38–40.

Bentler, P. M., & Bonnet, D. G. (1980). Significance tests and goodness of fit in the analysis of covariance structures. *Psychological Bulletin*, *88*, 588–606.

Boote, A. S. (1981). Reliability testing of psychographic scales: five-point or seven-point? Anchored or labeled? *Journal of Advertising Research*, *21*, 53–60.

Brown, G., Wilding, R.E., II, & Coulter, R.L. (1991). Customer evaluation of retail salespeople using the SOCO scale: A replication extension and application. *Journal of the Academy of Marketing Science*, *9*, 347–351.

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait–multimethod matrix. *Psychological Bulletin*, *56*, 81–105.

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

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

Cicchetti, D. V., Showalter, D., & Tyrer, P. J. (1985). The effect of number of rating scale categories on levels of inter-rater reliability: a Monte-Carlo investigation. *Applied Psychological Measurement*, *9*, 31–36.

Cohen, J. (1988). Statistical power analysis for the behavioral sciences, (2nd ed.). Hillsdale, NJ: Elrbaum.

Cohen, J. (1992). A power primer. *Psychological Bulletin*, *112*, 155–159.

Cox, E. P. (1980). The optimal number of response alternatives for a scale: a review. *Journal of Marketing Research*, *17*, 407–422.

Finn, R. H. (1972). Effects of some variations in rating scale characteristics on the means and reliabilities of ratings. *Educational and Psychological Measurement*, *34*, 885–892.

Garner, W. R. (1960). Rating scales, discriminability and information transmission. *Psychological Review*, *67*, 343–352.

Green, J. A. (1992). Testing whether correlation-matrices are different from each other. *Developmental Psychology*, *28*, 215–224.

Green, P. E., & Rao, V. R. (1970). Rating scales and information recovery: How many scales and response categories to use? *Journal of Marketing*, *34*, 33–39.

Hancock, G. R., & Klockars, A. J. (1991). The effect of scale manipulations on validity: targeting frequency rating scales for anticipated performance levels. *Applied Ergonomics*, *22*, 147–154.

Howell, D. C. (1992). *Statistical methods for psychology*. Boston, MA: Duxbury Press.
Jenkins, Jr., G. D., & Taber, T. D. (1977). A Monte-Carlo study of factors affecting three indices of composite scale reliability. *Journal of Applied Psychology*, *62*, 392–398.
Jones, R. R. (1968). Differences in response consistency and subjects' preferences for three personality inventory response formats. In *Proceedings of the 76th Annual Convention of the American Psychological Association* (pp. 247–248).
Komorita, S. S. (1963). Attitude content, intensity, and the neutral point on a Likert scale. *Journal of Social Psychology*, *61*, 327–334.
Lissitz, R. W., & Green, S. B. (1975). Effect of the number of scale points on reliability: a Monte-Carlo approach. *Journal of Applied Psychology*, *60*, 10–13.
Loken, B., Pirie, P., Virnig, K. A., Hinkle, R. L., & Salmon, C. T. (1987). The use of 0–10 scales in telephone surveys. *Journal of the Market Research Society*, *29* (3), 353–362.
Martin, W. S. (1973). The effects of scaling on the correlation coefficient: a test of validity. *Journal of Marketing Research*, *10*, 316–318.
Martin, W. S. (1978). Effects of scaling on the correlation coefficient: additional considerations. *Journal of Marketing Research*, *15*, 314–318.
Matell, M. S., & Jacoby, J. (1971). Is there an optimal number of alternatives for Likert scale items? Study 1: reliability and validity. *Educational and Psychological Measurement*, *31*, 657–674.
McKelvie, S. J. (1978). Graphic rating scales: How many categories? *British Journal of Psychology*, *69*, 185–202.
Messick, S. (1993). Validity. In R. L. Lin, *Educational measurement* (3rd ed.) (pp. 13–103). Phoenix, AZ: Oryx Press.
Miller, G. A. (1956). The magical number seven, plus or minus two: some limits on our capacity for processing information. *Psychological Review*, *63* (2), 81–97.
Neuman, L., & Neuman, Y. (1981). Comparison of six lengths of rating scales: students' attitude toward instruction. *Psychological Reports*, *48*, 399–404.
Nunnally, J. C. (1967). *Psychometric theory*. New York: McGraw-Hill.
Nunnally, J. C. (1970). *Introduction to psychological measurement*. New York: McGraw-Hill.
Oaster, T. R. F. (1989). Number of alternatives per choice point and stability of Likert-type scales. *Perceptual and Motor Skills*, *68*, 549–550.
Peabody, D. (1962). Two components in bipolar scales: direction and extremeness. *Psychological Review*, *69*, 65–73.
Peter, J. P. (1979). Reliability: A review of psychometric basics and recent marketing practices. *Journal of Marketing Research*, *16*, February, 6–17.
Ramsay, J. O. (1973). The effect of number of categories in rating scales on precision of estimation of scale values. *Psychometrika*, *38*, 513–533.
Remington, M., Tyrer, P. J., Newson-Smith, J., & Cicchetti, D. V. (1979). Comparative reliability of categorical and analogue rating scales in the assessment of psychiatric symptomatology. *Psychological Medicine*, *9*, 765–770.
Remmers, H. H., & Ewart, E. (1941). Reliability of multiple-choice measuring instruments as a function of the Spearman–Brown prophecy formula. *Journal of Educational Psychology*, *32*, 61–66.
Schutz, H. G., & Rucker, M. H. (1975). A comparison of variable configurations across scale lengths: an empirical study. *Educational and Psychological Measurement*, *35*, 319–324.
Shaw, M. E., & Wright, J. M. (1967). *Scales for the measurement of attitudes*. New York: McGraw-Hill.
Simon, H. A. (1974). How big is a chunk? *Science*, *183*, 482–488.
Symonds, P. M. (1924). On the loss of reliability in ratings due to coarseness of the scale. *Journal of Experimental Psychology*, *7*, 456–461.
Woodruff, D. J., & Feldt, L. S. (1986). Tests for equality of several alpha coefficients when their sample estimates are dependent. *Psychometrika*, *51*, 393–413.