Open access • Journal Article • DOI:10.1177/014662169401800101

A Comparison Between the Rating Scale Model and Dual Scaling for Likert Scales

Kwok-cheung Cheung, L. C. Mooi

Institutions: University of Macau, Nanyang Technological University

Published on: 01 Mar 1994 - Applied Psychological Measurement (Sage Publications)

Topics: Item response theory, Likert scale, Scale (ratio), Rating scale and Interval Scale

A Comparison Between the Rating Scale Model and Dual Scaling for Likert Scales

K. C. Cheung, University of Macau

L. C. Mooi, Nanyang Technological University, Republic of Singapore

The fundamental requirement of using Likert scales to measure affective behaviors—that all respondents must use the ordered response categories of all scale statements in the same way—is examined. Scaling problems arise when attitude statements within a Likert scale are unfolding preference data, and when the position and spacing of the ordered response categories are interpreted differently across scale statements and among respondents. The assignment of a neutral response category in the center of the response scale is questionable. These problems pertain to issues of the existence of an interval scale, the dimensionality of the trait, and patterns of item response functions. An attempt is made to resolve these problems using two contrasting scaling methods—item response theory modeling and dual scaling. The primary result is that conformity of responses to the item response theory model requirements and targeting of scale statements and their ordered response categories to the respondents in dual scaling are vital for a resolution of these problems. This study establishes the similarity of the two scaling methods. *Index terms: dual scaling, item response theory, jackknife technique, Likert scaling, Rasch model, rating scale model.*

In the late 1920s to early 1930s, Thurstone (1928) and Likert (1932) laid the foundation for modern psychological measurement of opinions, values, and attitudes. Likert proposed that a person's attitude toward an object of interest can be evaluated by a series of statements that can be judged as either favorable or unfavorable toward the target attitude object. Typically, each statement in a Likert scale is evaluated against a 5-point response scale of varying intensity of affective responses consisting of *strongly agree, agree, not sure, disagree,* and *strongly disagree.*

Although there are many variations of the Likert scale, Likert scales generally include an equal number of positively and negatively phrased statements, all of which employ the same response scale and are randomly distributed throughout the questionnaire. Numerical scores of 1 to 5 frequently are assigned to each statement in accord with the direction of the statement and the ordering of response categories. An attitude scale score then is obtained by aggregating these statement scores, typically by a simple summation. For the purpose of scale construction, Likert (1932) suggested that only statements with scores that are significantly correlated with the attitude scale score be included in the total scale. This statistical criterion is known as Likert's criterion of internal consistency. This criterion is the same as that of traditional test theory and is not sufficient to guarantee a truly unidimensional scale because two or more approximately equal but conceptually independent subsets of attitude statements belonging to different dimensions may be contained within the same scale (McIver & Carmines, 1981).

By examining the probability of responses along a presumed latent trait measuring an attitude, Thurstone (1959) differentiated between two types of attitude scales: (1) the maximum probability type in which the response value is the most discriminative at a certain location along the continuum of the trait and (2) the increasing probability type in which the response value varies monotonically along the continuum of the trait marked with category response thresholds. The

APPLIED PSYCHOLOGICAL MEASUREMENT
Vol. 18, No. 1, March 1994, pp. 1-13

0146-6216/94/010001-13$1.90

Likert scale is based on ordered response categories and is of the increasing probability type. Andrich (1989) indicated that Likert attitude data can be described by an unfolding model.

Modeling of Rating Scales

Modeling of Rating Scales Using Item Response Theory

Because of recent advances in psychometric theory, the fundamental requirement when using a Likert scale—that all respondents use the ordered response categories of all scale statements in the same way—now can be tested empirically. Likert scales require that the response categories must be equally-spaced, ordered response categories that can be structured on a unidimensional latent continuum. These categories should be applicable to all statements in the attitude scale so that consecutive integer values can be assigned to the ordered response categories.

This requirement is linked to issues concerning the number of response categories, the direction of the response scale, the inclusion of the Undecided category, and the position of a "middle/neutral" point on the response scale. Consequently, before a Likert scale is used, Likert's criterion of internal consistency should be evaluated. This involves evaluating the statistical fit of the responses to a model of item response behavior, such as the rating scale model (RSM). The RSM was developed as a Rasch logistic model within the family of item response theory (IRT) models.

Before the statements and the ordered response categories of a Likert scale can be calibrated for subsequent measurement purposes they should meet the assumptions of an appropriate IRT model (Cheung, 1990, 1991; Spearritt, 1982). Developed from the Rasch model of dichotomous responses (Rasch, 1960; Wright & Stone, 1979), both the RSM (Andrich, 1978; Masters, 1980; Mooi & Cheung, 1990; Wright & Masters, 1982) and the partial credit model (Andrich, 1988; Loh & Cheung, 1991; Masters, 1982; Masters & Wilson, 1989; Masters & Wright, 1984; Wright & Masters, 1982) are valuable in modeling ordered response categories onto a unidimensional continuum. The concept of an equal interval scale varies in these different modeling procedures (Andrich, 1988; Andrich & Schoubroeck, 1989; Linacre, 1990). The IRT definitions of equal-interval emphasize that the parameters in the IRT model are in an additive form so that their differences are comparable.

Andrich's (1978) RSM focuses on the transitions between adjacent ordered response categories by imposing a common response scale on all statements in the attitude scale. Each statement has a location estimate that is centered at the mean of the thresholds of its response categories, which are points on the trait continuum at which a respondent has an equal probability of choosing between adjacent categories. If the response thresholds are not correctly ordered along the trait continuum, the probabilities of endorsing some response categories are never higher than both of their adjacent categories. This is an indication that some categories are not functioning as intended by the Likert scale. The partial credit model (Andrich, 1978, 1988; Masters, 1982) relaxes the condition that each statement should conform to a common response scale. Douglas (1982) formulated a generic Rasch model that defines all possible variations of the Rasch family, in which the RSM and partial credit models are particular cases.

Andrich (1982) showed how IRT modeling recapitulates some key features of Likert scaling. Although Likert did not base his scaling procedure on an explicit response model, Likert (1932) empirically demonstrated the adequacy of integer scoring of the ordered response categories. Andrich (1982) observed that although the relationship between the total score of the attitude scale and the trait level estimate from Rasch analysis is nonlinear, there is a wide spectrum of the trait level within which this relationship is linear and perfect. This explains the success of integer scoring and that the total scale score approximates an interval measurement scale.

However, the inclusion of the *Not sure/*

Undecided category in the middle of the response scale as representing a neutral point threatens the assumption of unidimensionality. From the perspective of IRT modeling, this violation represents model misspecification and is likely to result in incorrectly ordered category thresholds.

The RSM

The RSM can be developed from the Rasch simple logistic model:

$$\frac{\pi_{ni1}}{\pi_{ni0} + \pi_{ni1}} = \frac{\exp(\beta_n - \delta_{i1})}{1 + \exp(\beta_n - \delta_{i1})}, \quad (1)$$

where π_{ni1} and π_{ni0} are the probabilities of person n with trait level β_n endorsing Category 1 (correct) and 0 (incorrect) of item i, respectively. For dichotomous data, δ_{i1}, the difficulty of item i, governs the probability of a response occurring in Category 1 rather than in Category 0 of item i. The left-hand side of Equation 1 is expressed in the form of a conditional probability involving two adjacent categories and, in this particular case, $\pi_{ni0} + \pi_{ni1} = 1$.

For polychotomous data, Equation 1 can be applied similarly to each pair of adjacent response categories and all properties of the Rasch model are preserved. For example, for a Likert scale of $m + 1$ categories (i.e., 0, 1, 2, ..., m),

$$\frac{\pi_{nix}}{\pi_{ni(x-1)} + \pi_{nix}} = \frac{\exp(\beta_n - \delta_{ix})}{1 + \exp(\beta_n - \delta_{ix})}, \quad (2)$$

where $x = 1, 2, \ldots, m$; and $\pi_{ni0} + \pi_{ni1} + \ldots + \pi_{nim} = 1$.

Furthermore, because in the RSM the same response scale is applicable to all item statements in the attitude scale, Equation 2 can be simplified further by proposing a single scale value δ_i for each item and a common set of transitional thresholds τ_x characterizing the functioning of the $m + 1$ response alternatives. Thus,

$$\delta_{ix} = \delta_i + \tau_x, \quad (3)$$

where $x = 1, 2, \ldots, m$.

The RSM implies that the persons, items, and categories (i.e., n, i, and x in Equations 2 and 3) are independent and that there are no interactions of any order (i.e., the relationships of persons, items, and categories in Equations 2 and 3 are additive). Moreover, the requirement of specific objectivity demands that the conditional probability involving two adjacent categories in Equation 2 is of the logistic form, which is invariant across persons, items, and categories. Therefore, the test of fit of the response data to the RSM would test the hypothesis that the Likert response categories are being used in the same way across all statements by the respondents. Unfortunately, any of these assumptions and requirements may fail empirically, and it is generally difficult to determine which is causing a failure of the model.

Dual Scaling

Dual scaling (Nishisato, 1980) is an alternative method for the analysis of Likert scales. Dual scaling allows the response data structure to be analyzed without relying on any prior assumption of unidimensionality, spacing or ordering of response categories, and the form of item response functions. The response categories comprising the Likert response scale are cross-tabulated with the statements to form a contingency table. Dual scaling seeks to quantify the row and column categories of the contingency table along parsimonious structural dimensions.

Dual scaling assigns weights (known as optimal weights) to the categories to maximize simultaneously the between-row and between-column sums of squares (SS) in relation to the total SS. [For details on how information in the contingency table is extracted along successive orthogonal structural dimensions in the context of a set of linear matrix equations see Nishisato (1980)]. Optimal weights are assigned to the row and column categories along these dimensions such that both between-row and between-column discriminations are maximized simultaneously. This sole criterion of dual scaling is Guttman's principle of internal consistency (for a formulation of dual scaling using this criterion, see Nishisato, pp. 21–27).

Dual scaling resembles a multidimensional

decomposition of data with the most informative structural dimension extracted first, then the second most informative dimension, and so on, until the information in the data is exhaustively extracted. Associated with each structural dimension is a statistic called "delta partial," which indicates the percentage of information in the table explained by that dimension. When there is a need to compare between structural dimensions, the optimal weights are further weighted in order to reflect their relative importance in extracting information. (For a discussion of the proper use of optimal weights when comparing the relative contribution of the structural dimensions, see Nishisato, 1980, p. 46.) The weighted optimal weights of a structural dimension can be plotted against each other to reveal the underlying structure of the contingency table. As such, dual scaling is model-laden although the maximization process is done without recourse to the order and distributional assumptions of the classifying categories (Nishisato, 1980, p. 68).

Purpose

This study was designed to compare the results of dual scaling and the RSM on a common dataset. The analysis was designed to evaluate both similarities and differences between these two approaches to the analysis of Likert scale data.

Method

Instrument and Calibration Sample

Respondents rated statements (items) on the Students' Liking for Computer-Related Activities (SLCA) scale using a 6-point ordered response Likert scale. The response categories were Dislike a Lot (labeled C1), Dislike (C2), Dislike a Little (C3), Like a Little (C4), Like (C5), and Like a Lot (C6). C7, which was not scored, was the nonresponse category. Respondents were instructed to not respond if they had no opinion to an item. The 15 items that comprise the SLCA are shown in Table 1. An integer scoring scheme was used.

The SLCA was administered to all Grade 12 females (N = 326) in the Arts, Commerce, and Science faculties of a junior college in Singapore (only eight respondents had studied computer programming). Cronbach α reliability was .96. This was high because of the central location of the item scale values in the response scale (item means ranged from 2.72 to 3.59; see Table 1) and the wide spread of the responses on each of the items in the attitude scale (item standard deviations ranged from 1.06 to 1.24).

Table 1
Means and Standard Deviations (SDs) for Students' Liking for Computer-Related Activities Scale Items (N = 326)

Item	Item Description	Mean	SD
1	Collecting computer brochures	3.01	1.14
2	Discussing computer software	3.35	1.16
3	Going to computer exhibitions	3.59	1.17
4	Reading books about computers	3.09	1.08
5	Taking part in computer programming competitions	2.72	1.11
6	Buying books on computer programming	3.07	1.14
7	Visiting computer installations	3.23	1.14
8	Reading computer magazines	3.05	1.06
9	Watching film shows featuring computers	3.15	1.12
10	Talking with friends about computers	3.13	1.11
11	Listening to friends talking about computers	3.26	1.14
12	Visiting software houses	3.37	1.16
13	Watching television programs on computers	3.46	1.23
14	Buying books on computers	2.96	1.10
15	Listening to talks on computers	3.15	1.24

The responses to the SLCA were analyzed separately using the RSM and dual scaling. For the initial analyses, each method used the information from all items (Solution 1). Then both scaling methods were compared using a dataset in which misfitting items, misfitting persons, and nonresponses were excluded ($N = 240$, Solution 2). In Solution 3, a constrained version of the dual scaling method was used to handle the incorrectly ordered response categories from Solution 2. In the third analysis, the standard errors (SEs) of optimal weights of C1 and C2 from Solution 2 were assessed using the jackknife technique.

Dual Scaling Analysis

The 326 responses to each of the n items ($n = 15$) for the $m + 1$ response categories ($m = 5$), including C7, were tabulated to form a two-way contingency table. It was assumed that the respondents used the $m + 1$ response categories in the same way across all n items. Optimal weights were calculated with the program DUAL3 (Nishisato & Nishisato, 1986) using the contingency/frequency table option, by maximizing the squared correlation ratio η^2—the ratio of between SS to total SS. These optimal weights (δ_i and τ_x for the items and response categories, respectively), after weighting to reflect the relative contribution of the dimensions, then were structured along the underlying dimensions. Although the same notation for the item and response category parameters used in the RSM were used here, their meanings should be evaluated and understood with reference to the scaling model. Specifically, when the response data conform perfectly to a Guttman scale, the second dimension obtained from dual scaling is often an "intensity" dimension, in the sense that it is just a quadratic function of the first dimension.

RSM Analysis

The RSM analysis was implemented by the computer program BIGSCALE (version 1.5; Wright, Linacre, & Schultz, 1989). Without dropping cases who selected C7 on any of the 15 items, the RSM could not be applied to the response data because of the unrealistic assumption that C7 is on the same unidimensional trait as that of C1–C6. Therefore, respondents who answered C7 were not used in this analysis. Misfitting respondents and items also were eliminated using the INFIT and OUTFIT statistics with the misfit criterion set at $|2|$. BIGSCALE provided scale values and the fit statistics of the fitting items, a map describing the response category probability functions of the items, and regions of most probable responses along the trait defined by the fitting items.

Results

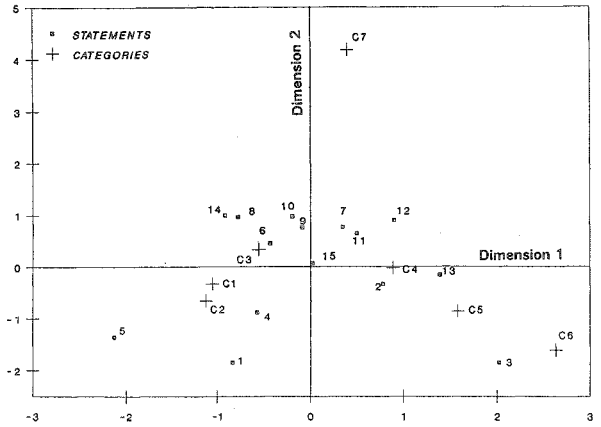
Dual Scaling

The dual scaling analysis accounted for 95% of the information in the data in three significant dimensions, with Dimensions 1–3 accounting for 75%, 14%, and 6%, respectively. The two-dimensional plots of optimal weights (weighted) for pairs of the three dimensions are displayed in Figure 1. The first dimension corresponded to the attitudinal continuum, along which items were arranged in order of affectivity. The difference in spacing between response categories C5 and C6 on Dimension 1 was moderate. [Note that whether the spacing between C6 and C5 was reliable or not remains an empirical question to be answered later due to the small percentage of responses (1.7%) used for calibrating C6]. The spacing on Dimension 1 was widest between C3 and C4, which was the region of neutral attitudes. By contrast, the spacing between C2 and C1 was too small to be regarded as discriminable. Whether a 5-point response scale, resulting from collapsing C1 and C2, would have been more suitable or not is discussed below in resolving a seeming paradox between the dual scaling and the RSM results.

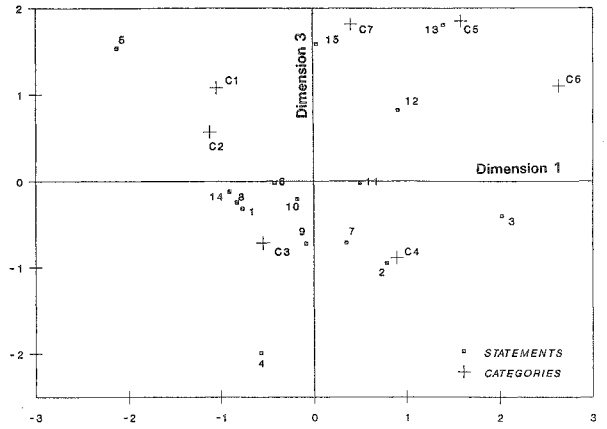
C7 was located between C3 and C4. The items on the lower end of Dimension 2 refer to activities that "are good to know" and information regarding them is "worth gathering" (e.g., Items 1, 3, 4, and 5; see Table 1). The upper end was

Figure 1
Two-Dimensional Plots of the Three Structural Dimensions From Dual Scaling (N = 326)

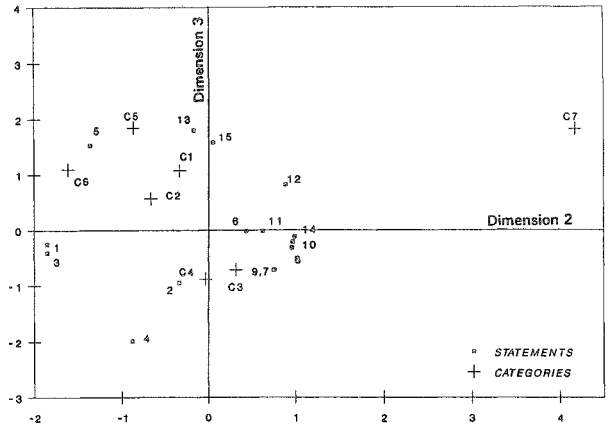
a. Dimensions 1 and 2



b. Dimensions 1 and 3



c. Dimensions 2 and 3



marked only by C7. Dimension 2 might not be interpreted as an intensity dimension because C7 rendered the response data unlikely to conform to a perfect Guttman scale.

Dimension 3 appeared to account for the item and response category interactions. The three orthogonal dimensions together accounted for approximately 95% of the information in the contingency table.

Rating Scale Analysis

C7 was not included in the RSM analysis because it was not on the same dimension as C1–C6. Among the 326 respondents, only 268 did not select C7 for any of the 15 items. Table 2 summarizes the results of the RSM analysis. Four items (Items 1, 5, 13, and 15) were classified as misfits. Item 8, which showed some signs of marginal misfit, was kept in the analysis because the negative sign for the INFIT and OUTFIT statistics indicated dependency in the data (see Wright, Linacre, & Schultz, 1989, p. 30 for a definition of the two fit statistics). In brief, INFIT is the information-weighted fit statistic, and OUTFIT is the outlier-sensitive fit statistic. Both are obtained when their mean-square fit statistics (MNSQ) are normalized; values substantially greater than 1 indicate "noise" in the data. Among the 268 respondents, 28 (10%) were classified as misfitting respondents; therefore, the calibration sample was further reduced to 240.

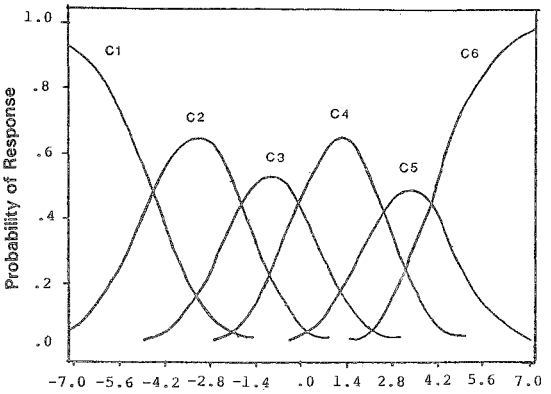
Table 2
Item Scale Values and Fit Statistics From the RSM (Solution 1) After Eliminating Misfitting Respondents and Items 1, 5, 13, and 15 (N = 240)

Item	δ_i	Error	MNSQ	INFIT	OUTFIT
14	.45	.09	.9	-1.1	-1.3
6	.34	.09	.9	-.6	-.7
4	.33	.09	.9	-.7	-.6
8	.24	.09	.8	-2.4	-2.4
10	.14	.09	.8	-1.8	-1.8
9	.12	.09	1.0	.1	.4
7	-.04	.09	1.1	1.4	1.4
2	-.18	.09	.9	-.7	-.8
11	-.24	.09	1.1	1.3	1.2
12	-.37	.09	1.0	-.5	-.4
3	-.79	.09	1.0	.6	.5

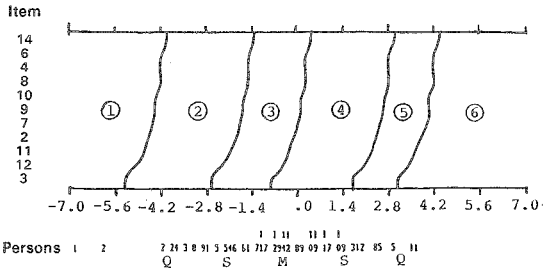
Figure 2a shows the probabilities of response to each of the six response categories for a given β relative to the scale value of each of the items (i.e., $\beta_n - \delta_i$). The intersections of the category probability functions are the set of transitional thresholds between adjacent response categories (i.e., $\tau_1, \tau_2, \ldots, \tau_5$). Figure 2b shows the regions of most probable responses of each of the items. The regions are depicted to show the relative position of the item scale value (i.e., δ_i) and the common set of transitional thresholds across all items (i.e., τ_x). The distribution of trait values of the respondents also is shown in Figure 2b.

Figure 2
Response Category Probability Functions and Regions of Most Probable Responses From the RSM (N = 240)

a. Response Category Probability Functions of Items



b. Regions of Most Probable Responses (Circled)



Comparison Between the Two Scaling Methods

Compared with other items, the misfitting items in the RSM (Items 1, 5, 13, and 15) tended to load on either or both of Dimensions 2 and 3 of the dual scaling (see Figure 1). These misfitting items were detected because of their violation

of the assumption of unidimensionality. Because Dimension 2 was essentially a nonresponse trait, the issue of proper treatment of a nonresponse is thus one of finding its causes, rather than determining a suitable location somewhere in the middle of the response scale in order to score the nonresponses on the trait that is to be measured.

The structuring of the items along dual scaling Dimension 1 was comparable in sequence to the item scale values obtained from the RSM (with the exception of Item 11, which showed some signs of instability when the misfitting items were present, and also the marginally misfitting Item 8; compare δ_i in Table 2 with the optimal weight of Solution 1 in Table 4). Using the present scoring methods, items with high scale values (i.e., the more affective items) in the RSM analysis were found to be associated in a reverse direction with the least affective end of the Likert response scale (i.e., the Dislike pole) in the dual scaling analysis, and vice versa. However, the direction of correspondence between the two scaling methods is actually a matter of convention. Remembering that the response categories can be viewed as signposts according to which students' responses are made, it is clear that items with higher affective values (i.e., those requiring more affective behaviors) are endorsed less readily at high levels of affectivity by an average student. Consequently, on average, students' responses to these more affective items are attracted toward the lower affective end of the Likert response scale.

Results of dual scaling and the RSM complement each other for an informed understanding of the issues of spacing and ordering of a response scale. The RSM assumes that the response categories are ordered and a violation of this assumption is regarded as model misspecification. Dual scaling treats the response categories as signposts that may be ordered by simultaneously ordering items with respect to a set of latent dimensions so that both response categories and attitude items are maximally discriminated. Thus, incorrectly ordered category thresholds resulting from the RSM and dislocated category optimal weights along dimensions from dual scaling are regarded as evidence of a poor response scale format which should be revised if a Likert format is what is intended. Without examining the SEs of the optimal weights, it could be concluded that the optimal weights of the response categories are sensibly ordered. The exception is on the Dislike pole of the response scale, which warrants a thorough examination by resolving a seeming paradox.

The Resolution of a Seeming Paradox

There is a seeming paradox in the spacing of the response categories between the two scaling methods—the spacing of the optimal weights of response categories in dual scaling and the spacing of the transitional thresholds between adjacent response categories in the RSM appear to be the inverse of each other. For example, the optimal weights of the two response categories C1 and C2 are not discriminable and even slightly disordered in dual scaling (see Figure 1), whereas there is a wide span of level of affectivity between C1 and C2 in the RSM (Regions 1 and 2, Figure 2b).

Viewed from the perspective of dual scaling, a possible explanation for category reversals (or overlap) of C1 and C2 is that the frequency pattern of C1 across the items is more similar than that of C2 to those of C3 to C6 (at least as it is captured on Dimension 1). This is because dual scaling is based on similarities of frequency patterns simultaneously among rows and columns. However, a fundamental requirement of any modeling procedure that seeks to examine underlying response structure is that there is systematic variation of response behaviors across items and response categories. The way the contingency table is input into the dual scaling analysis captures these consistent response behaviors despite the fact that the identity of the individual student has been lost. Consequently, relative consistency of responses is not mutually exclusive to test targeting, which is a crucial consideration in the RSM.

This paradox is thus resolved by considering that targeting of the affective level of items and the associated ordered response scale to the sample is vital for an adequate calibration of items and ordered response scale that define the trait to be measured (see the three considerations on targeting below). One immediate conclusion is that if a comparison is done simply for the purpose of selecting one scaling method as more correct in providing information on spacing and ordering of response categories, then the richness and meaningfulness of the information obtained from both methods is undermined.

A resolution of this paradox is illustrated by subjecting the responses to a reanalysis by dual scaling (Solution 2) using the same calibration sample (N = 240 respondents) as used in the RSM analysis. All nonresponses, misfitting respondents, and those who consistently endorsed C1 and C6 on all 11 items used in the RSM analysis were excluded. Table 3 shows the input data matrix in the form of a contingency table with the items arranged in order of scale values from the RSM. The underlying assumption is that respondents use the ordered response categories in the same way across the different items. This assumption in the dual scaling reanalysis was satisfied because it had been evaluated by the RSM analysis.

The results from a constrained version of dual scaling, which takes into account the incorrectly ordered response categories using the method of successive data modification (Nishisato, 1980, pp. 164–171) are presented in Table 4 (Solution 3). The unconstrained results of Dimension 1 shown in Figure 1 also are shown in Table 4 (Solution 1). SEs of the optimal weights using the jackknife procedure on 24 subsamples based on the class clusters of the junior college also are provided (see Keeves & Cheung, 1990, for a rationale of the jackknife procedure).

Table 3
Input Data Matrix for Dual Scaling Solutions 2 and 3 Based on N = 240 and 11 Statements

Item	C1	C2	C3	C4	C5	C6
14	15	56	86	69	11	3
6	13	59	80	68	16	4
4	13	53	84	74	15	1
8	6	56	94	64	19	1
10	10	51	81	79	17	2
9	10	51	78	84	13	4
7	12	46	71	90	15	6
2	10	41	76	85	24	4
11	10	42	67	91	29	1
12	6	43	69	89	28	5
3	8	33	54	97	40	8
Total	113	531	840	890	227	39

With reference to the results in Tables 2, 3, and 4 and Figures 1 and 2, there are three important considerations regarding the issue of test targeting.

Consideration 1. The respondents endorsed C2 more than C5 (531 versus 227 responses, see Table 3) across the 11 items. They endorsed C1 more than C6 (113 versus 39 responses), whereas the number of endorsements of the middle two categories, C3 and C4, was comparable (840 and 890, respectively). The distribution of the responses thus was skewed heavily toward the Dislike pole.

A large number of responses located at the lower end of the trait does not necessarily guarantee that this portion of the trait can be calibrated adequately. The reason is that the relative locations of the optimal weights of the items and response categories are all responsible for spanning the affectivity levels of the measured trait. Thus, respondents must be located at the lower region of the trait, and they must have an appropriate range of affective values to systematically discriminate between adjacent response categories across the 11 items, which have a range of affective scale values. In this sense, dual scaling, while attempting to maximally discriminate among the locations of items and categories along a linear dimension, is evaluating empirically the conformity of the items (with the response scale) and respondents to the measured trait. However, this is done without recourse to an explicit item response model, but rather in terms of response consistency to the response scale.

The success of this evaluation is reflected in

Table 4
Optimal Weights From Dual Scaling for Dimension 1
With Fitting Items Ordered by Their RSM Scale Values

	Solution 1	Solution 2	Solution 3	
Item/ Category	Optimal Weight	Optimal Weight	Optimal Weight	Jackknife SE
Fitting Items				
14	-.91	-1.17	-1.18	.05
6	-.43	-.78	-.78	.06
4	-.57	-.85	-.85	.05
8	-.77	-.91	-.90	.06
10	-.19	-.45	-.45	.06
9	-.09	-.42	-.42	.07
7	.34	.09	.09	.07
2	.78	.44	.44	.08
11	.49	.78	.78	.07
12	.90	.94	.94	.06
3	2.02	2.33	2.33	.06
Misfitting Items				
1	-.83			
5	-2.13			
13	1.39			
15	.02			
Response Categories				
C1	-1.05	-.79	-.86	.10
C2	-1.12	-.88	-.86	.03
C3	-.55	-.74	-.74	.03
C4	.89	.66	.66	.05
C5	1.58	2.20	2.21	.06
C6	2.63	2.32	2.32	.20
Nonresponse Category				
C7	.39			

the results of the RSM, which showed that along the affective continuum between -5.6 and -4.2 logits there was no respondent available to define accurately the threshold between C1 and C2 (Regions 1 and 2, Figure 2b). That this threshold still can be calibrated in the RSM is due to the three respondents whose levels of affectivity bracketed this threshold, enabling it to be determined by the implicit logistic model equation once the calibration sample was found to conform to the logistic modeling requirements. This view of targeting explains why the dual scaling results in Table 4 show that the two optimal weights of C1 and C2 could not be discriminated well by the large number of respondents located at the lower end of the trait who should have a greater probability of endorsing C2 rather than C1 according to the RSM (see C1 and C2, Figure 2a).

Consideration 2. Because more respondents endorsed C2 than C5 (see Table 3), the entire profile of the probability distribution of C2 was elevated; however, those of C5 decreased (see the functions for C2 and C5 in Figure 2a). The consequence was that there was a wider region of affectivity endorsing C2 and a narrower region endorsing C5 (Regions 2 and 5, Figure 2b). The spacing between C3 and C4 also was reduced (Region 3), whereas it was the reverse between C4 and C5 (Region 4). Thus, the issue of spacing between ordered response categories is fundamentally linked to a holistic consideration of the category response functions (i.e., the probability distribution of each of the response categories)

with specific reference to the calibrated sample. Interpreting equal-interval response scales as equal distances between adjacent transitional response thresholds ignores the interdependence of category response functions and the richness of information regarding the conformity of respondents and items to the measured trait (see Andrich, 1988, for a comparable view regarding the spacing of the thresholds).

Consideration 3. A consideration of the spacing of the optimal weights would not be complete without an examination of the SEs of their estimates (see Table 4). First, the optimal weights conform in sequence to the item scale values from the RSM when their SEs have been taken into account. The possible exception is the marginally misfitting Item 8 that may be excluded from the analysis. Second, after comparing the span of the trait obtained by the two scaling methods, the jackknife SEs of the optimal weights (ranging from .05–.08) appear comparable with the calibration errors of the items (of the order .09) from the RSM (Table 2). (The SEs of the optimal weights are expected to be larger than those obtained with the RSM if the nonresponses, misfitting persons, and items also are included in the dual scaling procedure.) Third, although the disordered response categories C1 and C2 were constrained to be equal, their SEs were not. This shows that C1 (SE = .10) invited a greater amount of measurement errors than C2 (SE = .03). Fourth, the apparent moderate span between C6 and C5 was not reliable, because of the small amount of data used for calibrating C6 (SE = .20).

Based on these three considerations, the seeming paradox regarding the spacing of the response categories obtained from both sets of scaling results can be resolved—the crux of the problem is a careful consideration of targeting the items and response scale to a distribution of respondents with consistent response behaviors across all items and categories of the attitude scale. For either scaling method, this is the fundamental requirement of using a Likert response scale.

Finally, the versatility of dual scaling is demonstrated by considering Solution 1 in Table 4. If items loading on Dimensions 2 and 3 are excluded and the reliability of optimal weights has been assessed and taken into account, dual scaling is as effective as the RSM. Dual scaling, however, can handle nonresponses easily and has very few prior model requirements. However, new computer programs for the RSM (e.g., TITAN; Adams & Khoo, 1991) are being developed to handle missing data routinely by leaving out those missing responses, but not the responses to other items a respondent has made.

By making full use of the information in the data as in Solution 1, dual scaling maximizes reproducibility of individual responses from the scaled dimensions that are intended outcomes of measurement (for a defense of the generalizability of scaling results see Nishishato, 1980, pp. 204–205). Similarly, in testing the fit of the IRT models, the basic ingredient is the quality of the reproducibility of the individual responses from the parameters estimated. The Rasch family models are the probabilistic versions of the Guttman structure—the response pattern can be reproduced completely given the total score. Consequently, although specific objectivity and reproducibility of individual responses are hallmarks of the two contrasting scaling methods they are actually functionally equivalent in the sense that both seek to fulfill essentially the same purpose of measurement; both appear to have accomplished this effectively, based on the present results.

Conclusions

This study showed that dual scaling was useful for analyzing Likert data because of its less stringent data and modeling requirements. Misfitting items from the rating scale analysis either loaded on some other latent dimensions irrelevant to the trait being measured or were measured less adequately onto a common trait than the items that defined it. The location of the item scale values and order of response categories were found to be comparable between the dual scaling and rating scale analysis. Through a resolu-

tion of a seeming paradox regarding the spacing of response categories between the two scaling methods, the issue of test targeting and conformity of responses was found to be of paramount importance in scaling items and response categories. In the Rasch family models, test targeting is done using conformity of responses to an item response model, whereas for dual scaling, it is ensured through systematic variations of consistent response behaviors across items and response categories using suitably-sized groups of respondents located in designated regions of the underlying trait. Relative consistency of responses is not mutually exclusive to targeting. Both are required in both models; otherwise, there is nothing systematic to model and there is not enough data at levels at which they are relevant.

The issue of spacing and ordering of categories of a response scale leads to information that may be obtained from a study of category response functions and/or the contingency frequency table of responses across items and response categories. As such, the similarity of the two different scaling methods is evident, although each has its own characteristic features and modeling requirements.

References

Adams, R. J., & Khoo, S. T. (1991). *TITAN: The interactive test analysis system.* Melbourne: Australian Council for Educational Research.

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

Andrich, D. (1982). Using latent trait measurement models to analyse attitudinal data: A synthesis of viewpoints. In D. Spearritt (Ed.), *The improvement of measurement in education and psychology: Contributions of latent trait theories* (pp. 89–126). Australia: Australian Council for Educational Research.

Andrich, D. (1988). A general form of Rasch's extended logistic model. *Applied Measurement in Education, 1,* 363–378.

Andrich, D. (1989). A probabilistic item response theory model for unfolding preference data. *Applied Psychological Measurement, 13,* 193–216.

Andrich, D., & Schoubroeck, L. (1989). The General Health Questionnaire: A psychometric analysis using latent trait theory. *Psychological Medicine, 19,* 469–485.

Cheung, K. C. (1990). Bamboo stems and pigtails: Some thoughts on meaningful measurement and testing in the classroom. In K. C. Cheung, W. K. Koh, K. C. Soh, & L. C. Mooi (Eds.), *Meaningful measurement and testing in the classroom using the Rasch model: Some exemplars* (pp. 1–6). Singapore: Institute of Education. (ERIC Document Reproduction Service No. ED 326 544)

Cheung, K. C. (1991). Climbing up the competence ladder: Some thoughts on meaningful assessment of problem-solving tasks in the classrooms. In K. C. Cheung, L. C. Mooi, & W. F. Loh (Eds.), *Meaningful assessment of problem-solving activities in the classroom: Some exemplars* (pp. 1–10). Singapore: National Institute of Education. (ERIC Document Reproduction Service No. ED 337 488)

Douglas, G. A. (1982). Conditional inference in a generic Rasch model. In D. Spearritt (Ed.), *The improvement of measurement in education and psychology: Contributions of latent trait theories* (pp. 129–157) Australia: Australia Council for Educational Research.

Keeves, J. P., & Cheung, K. C. (1990). Significance testing and related issues. In K. C. Cheung, J. P. Keeves, N. Sellin, & S. C. Tsoi (Eds.), The analysis of multilevel data in educational research: Studies of problems and their solutions. *International Journal of Educational Research, 14,* 299–306.

Likert, R. (1932). A technique for the measurement of attitudes. *Archives of Psychology, 140,* 1–54.

Linacre, J. M. (1990, April). *Modelling rating scales.* Paper presented at the annual meeting of the American Educational Research Association, Boston.

Loh, W. F., & Cheung, K. C. (1991). On meaningful measurement: Stages of lower secondary pupils' abilities in solving algebra word problems. In K. C. Cheung, L. C. Mooi, & W. F. Loh (Eds.), *Meaningful assessment of problem-solving activities in the classroom: Some exemplars* (pp. 29–49). Singapore: National Institute of Education. (ERIC Document Reproduction Service No. ED 337 488)

Masters, G. N. (1980). *A Rasch model for rating scales.* Unpublished doctoral dissertation, University of Chicago.

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika, 47,* 149–174.

Masters, G. N., & Wilson, M. (1989, March). *Understanding and using partial credit analysis: An IRT method for ordered response categories.* Notes prepared for the annual meeting of the American Educational Research Association, San Francisco.

Masters, G. N., & Wright, B. D. (1984). The essential process in a family of measurement models.

Psychometrika, 49, 529–544.
McIver, J. P., & Carmines, E. G. (1981). *Unidimensional scaling.* London: Sage Publications.
Mooi, L. C., & Cheung, K. C. (1990). On meaningful measurement: Junior college pupils' anxiety towards computer programming. *Journal of Educational Technology Systems, 19,* 327–343.
Nishisato, S. (1980). *Analysis of categorical data: Dual scaling and its applications.* Toronto: University of Toronto Press.
Nishisato, S., & Nishisato, I. (1986). *The DUAL3 statistical software series (IBM PC Version 3.10).* Islington, Ontario, Canada: Microstat.
Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests.* Copenhagen: Danmarks Paedogogiske Institut.
Spearritt, D. (Ed.). (1982). *The improvement of measurement in education and psychology: Contributions of latent trait theories.* Australia: Australia Council for Educational Research.
Thurstone, L. L. (1928). Attitudes can be measured. *American Journal of Sociology, 33,* 529–554.
Thurstone, L. L. (1959). *The measurement of values.* Chicago: University of Chicago Press.
Wright, B. D., Linacre, J. M., & Schultz, M. (1989). *A user's guide to BIGSCALE: Rasch model rating scale analysis computer program* (Version 1.5). Chicago: MESA Press.
Wright, B. D., & Masters, G. N. (1982). *Rating scale analysis.* Chicago: MESA Press.
Wright, B. D., & Stone, M. H. (1979). *Best test design: Rasch measurement.* Chicago: MESA Press.

Acknowledgments

The authors thank the two reviewers of this paper for generously sharing experiences and expertise.

Author's Address

Send requests for reprints or further information to K. C. Cheung, Faculty of Education, University of Macau, Caixa 3001, Macau.