DOCUMENT RESUME

ED 454 239 TM 032 837

AUTHOR Schulz, E. Matthew; Sun, Anji
TITLE Identifying Undifferentiating Response Sets and Assessing Their Effects on the Measurement of Items.
PUB DATE 2001-04-00
NOTE 43p.; Paper presented at the Annual Meeting of the American Educational Research Association (Seattle, WA, April 10-14, 2001).
PUB TYPE Reports - Research (143) -- Speeches/Meeting Papers (150)
EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS *College Students; High Schools; *Likert Scales; *Measurement Techniques; *Rating Scales; *Reliability; *Responses

ABSTRACT

Undifferentiating response sets, defined as "overuse" of any category of a Likert scale, were identified using a combination of simple criteria, such as whether a single-category response set involved more than four items, and statistical criteria based on D. Andrich's (1978) measurement model for Likert scales (the Rating Scale model). Data were from one section of the American College Testing Program's "Counseling for High Skills" survey for 10 colleges. Total counts across colleges for the 4 response sets were, respectively: 5,254; 4,757; 4,411; and 4,212. Undifferentiating response sets were strongly associated with statistically significant person misfit in Rating Scale model analyses. When persons with undifferentiating response sets were removed from the sample, the reliability of the item measures improved, and the rank order of the items became more internally consistent. It is concluded that applications of measurement theory can be useful in evaluating the quality of survey data. (Contains 4 figures, 9 tables, and 11 references.) (SLD)

Identifying Undifferentiating Response Sets and Assessing their Effects on the Measurement of Items

E. Matthew Schulz and Anji Sun

ACT, Inc.

Paper presented at the annual meeting of

The American Educational Research Association

April, 2001

Seattle, Washington

Abstract

Undifferentiating response sets, defined as "overuse" of any category of a Likert scale, were identified using a combination of simple criteria, such as whether a single-category response set involved more than four items, and statistical criteria based on Andrich's (1978) measurement model for Likert scales (the Rating Scale model). Undifferentiating response sets were strongly associated with statistically significant person misfit in Rating Scale model analyses. When persons with undifferentiating response sets were removed from the sample, the reliability of the item measures improved and the rank order of the items became more internally consistent. We conclude that applications of measurement theory can be useful in evaluating the quality of survey data.

In previous work (Sun and Schulz, 1999; Schulz and Sun, 2001) a latent variable model called the Rating Scale model (Andrich, 1978) was used with incomplete Likert data in order to control for the effects of a personal factor called "pleasability" on item ratings. Survey respondents' indicated their satisfaction with college services on a five-point Likert scale ranging from "very dissatisfied" to "very satisfied". Pleasability was defined as the tendency of a rater to use higher or lower ratings consistently across items. Because respondents were directed to rate only the services with which they had relevant experience, there were large amounts of missing data. Items were consequently exposed to different levels of pleasability. That is, groups responding to different items differed in average pleasability. In contrast to available case means analyses, Rating Scale model analyses controlled for the differential exposure of survey items to pleasability and yielded unbiased estimates of item performance.

A common form of the rating scale model is:

$$\ln\left(\frac{P_{nij}}{P_{nij-1}}\right) = \beta_n - \delta_i - \tau_j, \qquad (1)$$

where

ln means to take the natural log,

P_{nij} is the probability that person n chooses category j on item i,

P_{nij-1} is the probability that person n chooses category j-1 on item i,

β_n is the 'pleasability' of person n,

δ_i is the difficulty item i presents to feelings of satisfaction

τ_j is the relative difficulty of responding in category j or higher versus responding in category j-1; j=1,...,m-1; m is the number of response categories; the lowest category is coded 0.

Satisfaction is represented in this model by the *difference* between a person parameter (β_n) and an item parameter (δ_i), i.e., by (β_n - δ_i). *Ratings* of satisfaction are stochastically (probabilistically) related to this difference. The step parameters specify more fully the stochastic relationship between satisfaction and ratings of satisfaction. Due to the formulation of the model, more pleasable persons have higher values of β, and better performing items have lower values of δ.

In the present study, we explore the distinction between pleasability and an undifferentiating response set. We define an undifferentiating response set as "overuse of a single rating scale category." Response sets that consist of one category of the rating scale, such as all "1"s or all "5"s or all of any one category, such "3"s appear on their face to be undifferentiating. For example, ratings of all 5's contain no information about which item performed better than another. In contrast, pleasability involves a tendency to give higher or lower ratings, but not a tendency to restrict one's ratings to a single category.

Despite this distinction, there was no attempt in the previous work to statistically distinguish pleasability from an undifferentiating response set. For the purpose of measuring and comparing survey items, the distinction seemed of secondary importance. Both response tendencies confound or bias comparisons among items if they are distributed unevenly among the items.

Our interest in making a more formal study of the distinction between pleasability and undifferentiating response sets arose from the fact that a partial distinction along these lines was made automatically in the Rating Scale model analyses used in previous studies (Sun & Schulz, 1999; Sun & Schulz, 2000; Schulz & Sun, 2001). Respondents who assign all 1's or all 5's (i.e., either of the two extreme ends of the Likert scale) to the items they rate have inestimable parameters in the joint maximum likelihood estimation procedure used in standard software (e.g., Bigsteps, Facets, Mscale--all programs obtained through the University of Chicago) for Rating Scale analyses. These respondents, referred to as "extreme raters," comprised 2% of the respondents in one set of data involving a five-point Likert scale and 23 items (Sun & Schulz, 1999, 2001) and 10% in another set of data involving a four-point Likert scale and 10 items (Sun & Schulz, 2000).

Preliminary analyses with the latter set of data (Sun & Schulz, 2000) indicated that a substantial part of the effect of controlling for "pleasability," was in fact due to the elimination of extreme (and undifferentiating) raters in the Rating Scale analysis. When extreme raters were also excluded from the available case means analysis, the rank order of items by their available case mean rating was more consistent with their rank order by Rating Scale model analysis. The number of disagreements between the two methods about the relative performance of paired items fell from 36 to 28. Also, in cases of disagreement, the proportion of times the Rating Scale model correctly predicted which of the two items received the higher rating dropped from .531 when extreme raters were used to compute available case means, to .512 when they were not used.

These results suggested to us that available case means could yield an acceptable rank ordering of items with incomplete Likert data if undifferentiating raters were simply eliminated from the analysis. While we believe that a Rating Scale analysis is desirable on other grounds (Schulz & Sun, 2001), it might not be needed strictly for rank ordering items. A more thorough search for, and elimination of, undifferentiating raters could make it unnecessary to control for differential exposure of items to any remaining response tendencies, such as pleasability, left in the data. It is possible, for example, that items are differentially exposed to undifferentiating response sets, but not to response sets that reflect raters' levels of pleasability. Simple algorithms can be programmed to identify and eliminate any rater who uses just one category of the rating scale, including a middle or non-extreme category.

On the other hand, a Rating Scale model analysis could be a useful tool for identifying undifferentiating raters. Rating scale model's fit statistics include for each person and item a weighted mean square residual, *wmsr,* and a transformations of *wmsr* into an approximate t-statistic called *infit* (Wright & Masters, 1982; Smith, Schumaker & Bush, 1998). When data fit the model, *wmsr* has an expected value of 1; *infit* has an expected mean of 0 and variance 1. When a response pattern conforms to the model "too well" *wmsr* is less than 1, and infit is less than 0. These are referred to as cases of "overfit." Misfit also includes "underfit," which corresponds to *wmsr* greater than 1 and *infit* greater than 0. Undifferentiating response sets generally conform to the model too well in the sense that, because they display less random variability, or are less stochastic, than specified in the model, observed ratings tend to be too close on average to predicted ratings. A combination of practical and statistical criteria, such as *wmsr* < .6 and *infit* < -

2.0, are typically used to identify overfit to the Rating Scale model (Smith, et al., 1998; Schulz, 1990; Green, 1996b).

An advantage of using fit statistics to identify undifferentiating response sets is that a fit statistic systematically accounts for many factors simultaneously. The numeric value of a person fit statistic depends on the number of ratings in the pattern, the parameter estimates of the items rated, the parameter estimate (pleasability) of the rater, and the specific ratings given to the items. It would be difficult to account for these factors systematically and simultaneously with simple algorithms. Not all raters who use one category exclusively, or even *almost* exclusively, are necessarily undifferentiating raters. In particular, as the number of items rated by a respondent decreases, it is difficult to judge whether the use of just one category of the Likert scale reflects failure to differentiate the items, or just chance consistency in the experience of satisfaction across items. A standardized fit statistic (*infit*) combined with a practical criterion (*wmsr*) can be helpful in classifying such a response set as undifferentiating or not.

Disadvantages of fit statistics include their dependence on sample size and their lack of specificity. Single-category response sets involving a large number of items might be undifferentiating on their face, but might not yield statistically significant misfit. Response patterns associated with statistically significant misfit might nevertheless represent valid, differentiating information about the items. Not all response patterns that have a small (< .6) mean squared residual (*wmsr*) or negative infit (< -2), for example, will necessarily appear to be undifferentiating on their face (i.e., by inspection and informal judgement.)

The approach in this study is to assess the relationship between Rating Scale model fit statistics and simple (algorithmic and judgmental) criteria for undifferentiating response sets, and to use both to identify and eliminate undifferentiating raters. Simple criteria will include the number of items rated and whether a single category is used exclusively (SC for single category) or almost exclusively (ASC for almost single-category. We define ASC response patterns as those in which all but one rating is the same. These would include, for example, a respondent who rated a total of 9 items and who was "very satisfied" with all but one. We will not further differentiate ASC patterns by the specific rating given to the exception. The number of items rated is an important simple criterion for identifying undifferentiating response sets. An SC response set requires at least two items and ASC at least three, by definition.

For the purpose of implementing simple criteria for undifferentiating response sets as well as for assessing the association of Rating Scale model fit statistics with simple criteria, it makes sense to classify SC and ASC response sets by the number of items rated. If SC response patterns involving as few as 2 items, and ASC reponse patterns involving as few as 4 items were considered as possibly being undifferentiating response sets, and there were L total items in the survey, there could be up to L-1 SC groups and L-3 ASC groups for each category of the rating scale. Each SC group would contain persons who rated the same number of items and used the same Likert scale category exclusively. Each ASC group contains persons who rated the same number of items and used the same Likert scale category for all but one of the items.

We expect the fit statistics for persons classified into these groups to become more extreme as the number of items rated increases. We expect that only a minority of

SC and ASC patterns involving the fewest number of items will be flagged as misfitting the Rating Scale model (*wmsr* < .6 and *infit* < -2). We also expect that a majority of persons within SC groups involving large numbers of items, e.g., L, L-1, or L-2 items, will misfit the Rating Scale Model. A limitation of this investigation is that fit statistics are not available for extreme raters (raters not measured in an RSM analysis). This means that we can study the relationship between fit statistics and SC patterns involving only non-extreme categories. However, we can study the relationship between fit statistics and ASC patterns involving any category of the Likert scale.

In order to assess the effects of eliminating undifferentiating raters, four sets of data will be created. The first set (Set 1) will consist of all available data. Subsequent sets will be correspond to increasingly more inclusive criteria for classifying response patterns as undifferentiating. Eliminating extreme raters from Set 1 will create set 2. Set 3 will be created by eliminating from Set 2 SC groups in which 1) 4 or more items are rated and 2) all persons within the group misfit. Set 4 will be created by eliminating from Set 3, only the misfitting persons within SC groups rating 4 or more items and ASC groups rating 5 or more items. Set 4 will contain fewer SC patterns than Set 3 only if there are SC groups in which not all persons misfit.

Undifferentiating response patterns should either have no effect on, or should decrease the reliability of item measurement. Reliability is the empirical or theoretical correlation between two independent measures of the same items. The methods by which reliability will be estimated are presented in the methods section. The usual effect of eliminating raters would be to decrease reliability since reliability decreases with sample size. However, if the raters were not helping to differentiate items, or were contributing

bias or noise to item measurement, reliability might increase. Two measurement methods will be used: available case means (ACM) and Rating Scale model (RSM). Within each method, we expect the reliability coefficient to remain the same or to increase across the successive sets defined above (Sets 1 through 4). Reliability estimates will not be compared across methods because the procedures for estimating reliability differ, and involve different assumptions, across methods.

Undifferentiating response sets should have only slight effects on the reliability of person measurement. Green (1996b) reported that the reliability of person measurement increased slightly when misfitting persons were dropped. However, underfitting as well as overfitting persons were dropped in that study. In the present study, we drop only overfitting persons who display undifferentiating response sets. Overuse of more extreme categories leads to more extreme levels of the measured trait (pleasability), greater variance of the distribution of person measures, and hence inflated reliability. By the same line of reasoning, overuse of a middle category would suppress reliability. The net effect on reliability will depend on whether undifferentiating response sets involve use of more extreme or middle categories.

Internal order consistency (IOC) rates will also be computed for each method (ACM and RSM) and set of data. IOC rates cannot be estimated for the RSM-by-Set 1 combination, however, because extreme raters are automatically excluded from RSM measurement. IOC is the degree to which items with higher marginal rank order are more likely to receive the higher rating with multiple items are rated by the same rater. Various procedures have been used to estimate the IOC of a given rank ordering

(Johnson, 1997; Schulz and Sun, 2001). Methods described in the methods section, and used by Schulz and Sun (2001), will be used in the present study.

Within each method, (ACM and RSM) we expect IOC rates to increase as the criteria for undifferentiating raters becomes more comprehensive. This expectation follows from the notion that undifferentiating raters do not contribute reliable and consistent information about differences between items. We expect IOC rates to increase most from Set 1 to Set 2 (a finding that can only be seen with ACM rank orderings.)

In order to assess the effect of controlling for response sets that remain in data after eliminating undifferentiating raters, internal order consistency (IOC) rates will be compared across measurement methods (ACM versus RSM). A rating scale analysis, but not available case means, controls for differential exposure of items to response sets of the type defined as pleasability (Schulz & Sun, 2001), but which may also be undifferentiating to some degree. If items are differentially exposed to remaining response sets, IOC rates should be higher for RSM than for ACM (Schulz and Sun, 2001).

The IOC rates of RSM and ACM will be compared through the RSM conditional internal order consistency (RSM CIOC). CIOC is computed conditionally on disagreement between two alternative rank orderings about the relative performance of items. The disagreement is detected in pairwise fashion, by considering the relative performance of two items in each ranking. For N items, there are N-choose-two possible cases of disagreement (if the number of items is even). In any case of disagreement, one finds all the raters who rated both items, and gave each a different rating. A given ranking's CIOC is the proportion of times it correctly predicted which item received the

higher rating. This rate may be computed by item pair, by item, or as an overall (all cases) value. The ranking with CIOC > .5 is the better ranking.

Methods

Data

The data was taken from one section of the ACT *Counseling for High Skills* (CHS) survey. The CHS survey is administered to students enrolled in post-secondary career/technical programs at colleges in the United States. The section used for this study contains ten items, shown in Table 1. They represent college services or facilities and include, for example, "academic counseling", "personal counseling", "job placement", and "designated study areas". Ratings are on a four-point Likert scale ranging from "poor" (=1) to very good (=4). A fifth category, "unable to evaluate", was provided and was treated as missing data.

Data from ten colleges were used for this study. The colleges were selected on the basis of sample size (the largest available). There was no requirement for selection other than sample size. No attempt was made to control for characteristics of the colleges such as public/private, four-year/two-year, affiliation, location, enrollment, etc. Sample sizes by college are shown in Table 2. Set 1 includes any person who rated at least one item. Set 1 sample sizes ranged from 344 in School 7 to 961 (School 1).

There was a significant amount of missing data because respondents rated only the items with which they had relevant experience. In Set 1, the average number of items rated per person, within colleges, ranged from 5.2 in College 7 to 7.1 in College 2. These figures mean that 29% to 48% of the ratings were missing, depending on the college. Within school sample size per item ranged from 58 (Item 5) to 941 (Item 3). These items

accounted for the lowest and highest percentage response rates (14% and 98% respectively). Across schools, the response rate was highest to Item 3 (Course scheduling and registration) and lowest to Items 5 (Housing assistance) and 9 (Child care).

The sample was edited based on undifferentiating response sets as described above. Table 2 shows sample sizes by college for Sets 1 through 4. Total counts across colleges for the 4 sets were, respectively, 5254, 4757, 4411, and 4212. Differences between these numbers show the number of raters eliminated according to increasingly inclusive criteria for undifferentiating response sets. For example, 497 raters (5254 minus 4757) were extreme raters. More detailed information about these and other undifferentiating raters, including the association between simple criteria and fit to the Rating Scale model, is presented in the results section.

Rating Scale Analyses

Rating scale analyses were performed using the Bigsteps computer program (Wright and Linacre, 1991). Three analyses were performed separately by school. These analyses yielded three sets of item parameter estimates based separately on persons in Sets 2 through 4. Fit statistics (*wmsr* and *infit*) of persons in Set 2 were tabulated by group of SC and ASC raters. Type of pattern, Likert scale category, and number of items rated, as described above defined the groups.

The reliability of item parameter estimates in Rating Scale analyses was computed as one minus the ratio of mean squared measurement error to the variance of the item parameter estimates. The measurement error of each item parameter is estimated

routinely in a Rating Scale analysis by the Fisher information function. Reliability of person measurement was computed in the same fashion.

Available case means analysis

An item's available case mean is the simple arithmetic average rating assigned to the item. The reliability of the available case mean was estimated by a split-half technique. For each set of data (Sets 1 through 4) raters within a school were randomly assigned to one of two split-half groups, and mean item ratings were computed for each group. The split-half correlation between the mean item ratings was corrected for attenuation using the Spearman-Brown formula.

Internal Order Consistency (IOC) and Conditional IOC

IOC and CIOC were estimated using tabulation methods (Schulz and Sun, 2001). In each set of data within a school, we searched for *pairs* of ratings involving 1) the same rater, 2) different items (two different items), 2) different ratings (the items did not receive the same rating). Let *TOT1* represent the total number of such finds in a given set of data. For each find, a marginal ranking (ACM and/or RSM) was consulted to determine whether the items were in the same order as their ratings. If yes, a '*hit*,' was recorded. If the items had a tied ranking, a '*tie*' was recorded. Different rankings derived from the same data could have different totals for *hits* and *ties*, but *TOT1* is the same. The IOC rate for a given ranking is (*hits* + *ties*/2)/*TOT1*.

Conditional internal order consistency (CIOC) is computed using only cases of disagreement between two alternative rank orderings. We begin by searching for pairs of items in which one item in the pair is higher in one ranking but lower in the other. We find all such pairs of items. For each pair, we search for raters who assigned different

ratings to the items. Let TOT2 represent the total number of such raters summed over the total number of item pairs for which disagreement in rank was found. Hits are defined as above, but only one of the rankings can score a hit because only cases of disagreement are considered. The RSM CIOC is the total number of hits attributed to the RSM ranking divided by *TOT2*. The ACM CIOC is one minus the RSM CIOC. The method whose CIOC rate is above 0.5, is the better method.

Results

The frequency of undifferentiating response sets using simple criteria, and their association with Rating Scale model fit statistics, is shown in Tables 3 and 4. Table 3 includes persons who assigned all 2s (average) or all 3s (good) to the items they rated. These are non-extreme, SC response sets. Extreme SC response sets are not included because no fit statistics were estimated for these. Table 4 includes persons with ASC response sets involving any of the four Likert scale categories. In both tables, the response sets are grouped into rows by the total number of items rated.

SC patterns are far more likely than ASC patterns to misfit the Rating Scale model. The *wmsr* is extremely small (< .6) for all SC patterns, regardless of number of items rated, but for ASC patterns, *wmsr* is more variable and has an average value not too far from its expected value (1). The maximum *wmsr* for any SC pattern was .19 (for a person who rated seven items). Average *wmsr* for ASC patterns by number of items rated ranged from 1.03 (for seven items rated) to .8 (for six items rated).

Reflecting the effect of sample size on the power of a normalized fit statistic, the *infit* of SC and ASC response patterns to the Rating Scale model generally becomes worse as the total number of items rated increases. With an expected value of 0 under the

hypothesis of model fit, mean *infit* by SC group decreased from -2.16 with just two items rated to -3.70 with eight items rated, and remained near -3.7 for nine and ten items rated. For ASC patterns, *infit* decreased from -.36 with four items rated to -1.07 with ten items rated.

All raters within SC groups of 5 or more items misfit the Rating Scale model (Table 3). The maximum *wmsr* and *infit* values within these groups are less than 0.6 and -2.00 respectively. The total number of persons within these groups, 346, accounts for the difference between Set 2 and Set 3 sample sizes (Table 2). That is, these were the raters that were dropped from Set 2 in order to form Set 3. The criteria for Set 4 was to delete from Set 3, persons in the remaining groups (rows) of Tables 3 and 4 who misfit the Rating Scale model. The difference between sample sizes of Set 3 and Set 4 is 199 (4411 minus 4212), which is about twenty percent of the raters in these groups.

The raters in Tables 3 and 4 plus extreme raters, represent a large proportion of all raters. There were 497 extreme raters, 666 raters in Table 3, and 752 raters in Table 4. The total, 1915, is 36% of the grand total (5254 in Set 1). The extreme raters alone comprise 9.5% of the total.

It might be of interest to note that the majority of undifferentiating raters used the higher end of the Likert Scale. 444 of the 497 extreme raters used category 4 (very good). 488 of the 666 SC response patterns in Table 3 used category 3 (good). 656 of the 752 ASC response sets in Table 4 used category 3 or 4 predominantly.

IOC rates by college, data set, and method of ranking are shown in Table 5. Overall IOC rates (averaged across colleges) by data set and method of ranking are shown in the last row. For ranking items by available case means (ACM), IOC rates

were highest for Set 2 (.62705) and lowest for Set 1 (.62499). For ranking items by Rating Scale model analysis (RSM), overall IOC rates were highest for Set 3 (.62937), and lowest for Set 4 (.62782), but the Set 2 overall value was close to that for Set 2 (.62934). The lowest overall IOC rate for the Rating Scale model (.62782) was higher than the highest overall IOC rate for the ACM (.62705).

RSM conditional IOC rates in Table 5 show more clearly that, for any set of data where direct comparison was possible (Sets 2 through 4), the RSM ranking was more internally consistent than the ACM ranking. All RSM conditional IOC rates in the last row (overall) are above 0.5. RSM displayed the largest advantage (RSM CIOC = .52644) when the criteria for eliminating undifferentiating raters was most comprehensive (Set 4), and the least advantage (.51713) when the criteria was least comprehensive (Set 2, with elimination of only extreme raters.)

With few exceptions, within-school, within-Set results are consistent with overall results. Within schools and sets, RSM CIOC values were greater than 0.5 in 21 of 30 comparisons and were less than 0.5 in only four. 0.5 values mean that the ACM and RSM rankings did not differ. For ACM, Set 2 IOC values were equal to or higher than their Set 1 counterparts in nine out of ten schools; Set 4 IOC values were lower than those for Set 2 and Set 3 in six of ten schools. For RSM, Set 4 IOC values were lower than those for Set 2 and Set 3 in seven of ten schools.

Table 6 shows reliability estimates by measurement method (ACM and RSM), data set, and school. The last row shows overall (averaged across colleges) reliabilities by method and data set. In both methods, deletion of undifferentiating response sets had only slight effects on reliability. Up to Set 3, reliability slightly improved. With ACM,

reliability increased form .935 (Set 1) to .943 (Set 3). With RSM, reliability increased from .938 (Set 2) to (.948) Set 3. In both methods, reliability for Set 4 was lower than for Set 3 (.939 versus .943 for ACM; .945 versus .948 for RSM).

With few exceptions, within-school results were consistent with these trends. For ACM, Set 2 reliability was equal to or higher than Set 1 reliability in eight of ten schools. For RSM, Set 3 reliability exceeded Set 2 reliability in five of ten schools and was equal to Set 2 reliability in the remainder.

School 7 was a notable exception. Eliminating extreme raters substantially decreased ACM reliability (from .85 with Set 1 to .76 with Set 2). This result may be partly due to the fact that sample sizes for measuring the items in School 7 were small. The number of raters in School 7 was the smallest of any school (344 Set 1). The average number of items rated per person in School 7 (5.2 in Set 1) was the lowest of any school. Further reductions in sample size by eliminating extreme raters may have had overwhelming effects on the reliability of item measurement and on the error with which reliability is estimated.

We also note in this connection that the RSM CIOC value for Set 2 in School 7 is one of the highest in Table 5 (.56129). This value means that the items in School 7 were still differentially exposed to pleasability (or to remaining undifferentiating response sets) even after extreme raters were dropped. This result, and the decrease in ACM reliability from Set 1 to Set 2, suggests that differential exposure to extreme response patterns contributed a bias to item measures that was consistent with the bias that differential exposure to the remaining response sets were contributing to item measures. With ACM, then, dropping extreme raters had the sole effect of reducing sample size.

The reliability of person measurement decreased when undifferentiating response sets were dropped. The last row of Table 7 shows that average (across colleges) reliability coefficients of RSM measures from Set 2, Set 3, and Set 4 data were, respectively, .72, .69, and .68. Results within every school were consistent with these trends, except that in some cases, there was no difference between Set 3 and Set 4 reliability coefficients.

Discussion

The results of this study show that undifferentiating response sets can be distinguished from pleasability through a combination of simple algorithmic criteria and Rating Scale model fit statistics. Pleasability is defined as the trait that is still measured after deleting undifferentiating response sets. The reliability of person measurement was .72 on average with all data (Set 1), and .68, on average when undifferentiating response sets were removed using the most comprehensive criteria of simple algorithms and fit statistics (Set 4). Thus, considerable individual differences in the measured trait remained after removing undifferentiating response sets.

It is unclear, however, whether fit statistics are indispensable for identifying response sets whose removal improves the measurement of items. In defining Set 3 data, fit statistics were used to decide which SC groups should be dropped. All the persons within the dropped groups were flagged for misfit. This result certainly shows that fit statistics are strongly associated with simple criteria for undifferentiating response sets. However, the response sets of these groups might well have been judged by inspection to be undifferentiating. The quality of item measurement with Set 4 data, which was defined by more extensive and indispensable reliance on fit statistics, was generally not

as good as that with Set 3 data. In terms of the reliability and internal order consistency of item measures, Set 3 was better than Set 4 using both ACM and RSM measurement.

Our results suggest that SC response sets involving any category of a Likert Scale, not just extreme categories, can be profitably dropped for the purpose of measuring differences between items. This conclusion is based on the criteria for defining Set 3 and on the reliability and IOC of item measurement associated with this set of data, in comparison to other sets, using both ACM and RSM measurement. For Set 3, we removed SC sets involving as few as four items. We do not know whether a higher number, such as five or six items, would have been a better choice for our data or which choice would be best for any other set of data. Our results for Set 4 show that one can go too far--criteria can become too comprehensive.

Key aspects of our results might depend on specific characteristics of our data. Our Likert scale did not have a middle category, and there were only ten items. As a percentage of the total number of raters (5254), there were many extreme raters (9.5%), nonextreme SC raters (13%) and ASC raters (14%) in our study. More items in the survey might have decreased the number of SC and ASC patterns. A middle category might have decreased the number of SC response sets involving categories adjacent to the middle (categories 2 and 3 in our study). Only 2% of raters were extreme in a study involving more items (23 items) and a five-point Likert scale (Sun & Schulz, 1999).

Subsequent results might also depend on the number of items and Likert scale. While a middle category might have decreased the overall number of SC response sets, it might also have accounted for a very large, if not the largest, proportion of SC sets.

Deleting SC patterns might then have had the overall effect of increasing the reliability of person measurement.

Deleting undifferentiating response sets had only slight effects on internal order consistency. IOC was only slightly affected, primarily because there was little or no change in the rank order of the items. With more items, there might have been changes in rank order.

IOC and CIOC rates are crude, but simple methods for assessing difference in rank order. IOC rates for rank orders derived from the same data are typically extremely close because the rank orders may not differ, or may differ only for items whose "true" performance levels or ranks, are very similar. CIOC rates will not deviate very far from 0.5.

Since there are no known formal tests of statistical inference that can be performed on IOC and CIOC values, general conclusions must be based on informal considerations such as the consistency of findings across schools. Results within schools consistently favored Set 3 as the best set for measuring differences between items and the RSM as the better method of measurement. Effects on the reliability of item measurement were also small, and tests of significance are problematic for comparing indices derived from the same or overlapping sets of data. However, again, results were consistent across schools in pointing to Set 3 as the best set of data for measuring differences between items.

Our results do not support strong criticisms of analyses that include undifferentiating raters or even strong criticisms of analyses that analyze incomplete Likert data using conventional approaches, such as available case means. However, they

do support the conclusion that deletion of undifferentiating raters and use of the Rating Scale model can improve the measurement of items with incomplete Likert data.

More generally, we believe this study provides support for Green's (1996a) theme that applications of measurement theory can be useful in the evaluation of survey data quality (Green, 1996a). The use of measurement theory can lead to useful insights or conceptualizations, such as the distinction between pleasability and satisfaction, and the distinction between pleasability and undifferentiating response sets. These conceptualizations, supported by research, can lead to more general improvements, such as improvements in the design of surveys, and interpretations of results, if not to substantial improvements in simple outcomes, such as the rank order of items. For example, the results of this study may lead to the addition of a middle category to the Likert scale in this study. This modification would be for the purpose of improving the measurement of items, and would not necessarily improve the measurement of the raters.

References

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

Green, K. E. (1996a). Applications of the Rasch model to evaluation of survey data quality. In M.T. Braverman & J.K. Slater (Eds.), *Advances in Survey Research* (pp 81-92). Number 70. San Francisco: Jossey-Bass.

Green, K. E. (1996b). *The Use of Person Fit Statistics in Mail Surveys.* Paper presented at the annual meeting of the American Educational Research Association. New York, NY.

Johnson, V. E. (1997). An alternative to traditional GPA for evaluating student performance. *Statistical Science*, *12*, 251-278.

Smith, R. M., Schumaker, R. E. & Bush, M. J. (1998). Using item mean squares to evaluate fit to the Rasch model. *Journal of Outcome Measurement*, *2*, 66-78.

Schulz, E. M. (1990, Winter). Functional assessment of fit of data to the Rasch model. *Rasch Measurement SIG Newsletter*, *3* (4), 7-9.

Schulz, E. M. & Sun, A. (2001). *Controlling for rater effects when comparing survey items with incomplete Likert data.* ACT Research Report 2001-2, Iowa City, IA: ACT, Inc.

Sun, A., & Schulz, E. M. (2000). *A rating scale model procedure for comparing institutions with incomplete Likert data.* Paper presented at the annual meeting of the American Educational Research Association. New Orleans, LA.

Sun, A., & Schulz, E. M. (1999). *Rank ordering and comparing survey items using an IRT Rating Scale Model.* Paper presented at the annual meeting of the American Educational Research Association. Montreal, Canada.

Wright, B. D. & Linacre, J. M. (1991). *Bigsteps.* A Rasch-model computer program. Chicago: MESA Press.

Wright, B. D. & Masters, G. N. (1982). *Rating Scale Analysis.* Chicago: MESA Press.

Table 1.
Counseling for High Skills Survey Items Used in the Study

Item	Content
1	*Counseling (academic)*
2	*Counseling (personal)*
3	*Course scheduling and registration*
4	*Financial aid (grants, loans, etc.)*
5	*Housing assistance*
6	*Part-time job placement while enrolled as a student*
7	*Job placement (career-related)*
8	*Academic support (such as tutoring, study skills)*
9	*Availability of childcare*
10	*Designated study areas*

Note.
Ratings for these items are coded as 1=poor, 2=average, 3=good, and 4=very good. The response of "unable to evaluate" is coded as a missing value.

Table 2. Sample Sizes by College and Set

College	Set 1 Respondents who rated at least one item	Set 2 RSM-Measured (non-extreme) Raters	Set 3 No SC Patterns of more than 4 items.	Set 4 No misfitting SC or ASC patterns
1	961	862	792	745
2	484	459	427	408
3	414	379	350	337
4	374	324	303	292
5	837	771	704	669
6	398	356	342	329
7	344	311	281	269
8	556	512	477	456
9	402	354	336	319
10	484	429	399	379
Totals	5254	4757	4411	4212

Table 3. Fit Statistics for Non-extreme, Single-Category (SC) Groups

Number of Items Rated	Group Size[1]	Weighted mean squared residual (wmsr)[2]			Standardized wmsr (infit)[3]		
		Min.	Avg.	Max.	Min.	Avg.	Max.
10	198	0.04	0.1	0.17	-4.99	-3.68	-2.97
9	16	0.04	0.08	0.16	-4.64	-3.6	-2.88
8	17	0.02	0.06	0.15	-5.05	-3.7	-2.62
7	20	0.01	0.05	0.19	-4.9	-3.65	-2.54
6	34	0.01	0.04	0.17	-4.75	-3.43	-2.15
5	61	0	0.03	0.18	-4.35	-3.25	-2.03
4	90	0	0.02	0.14	-4.07	-3.03	-1.81
3	111	0	0.02	0.2	-3.72	-2.65	-1.26
2	119	0	0.01	0.12	-3.18	-2.16	-1.19
Total	666						

Note. [1]Include raters who used all 2s or 3s (non-extreme single-category categories)
[2]Expected wmsr = 1
[3]Expected infit mean = 0, standard deviation = 1

Table 4. Fit Statistics for Almost Single-Category (ASC) Groups

Number of Items Rated	Group Size[1]	Weighted mean squared residual (wmsr)[2]			Standardized wmsr (infit)[3]		
		Min.	Avg.	Max.	Min.	Avg.	Max.
10	72	0.06	0.86	3.9	-4.07	-1.07	4.1
9	32	0.06	0.93	3.51	-4.29	-0.7	3.34
8	45	0.14	0.93	3.55	-3.12	-0.53	3.49
7	64	0.06	1.03	3.8	-3.11	-0.19	3.72
6	144	0.06	0.8	3.96	-3.06	-0.58	3.67
5	163	0.08	0.85	4.2	-2.61	-0.36	3.53
4	232	0.07	0.81	5.54	-2.41	-0.36	3.65
Total	752						

Note. [1]Include raters who used same category for all but one items rated
[2]Expected wmsr = 1
[3]Expected infit mean = 0, standard deviation = 1

Table 5. Internal Order Consistency and Conditional Internal Order Consistency

	Internal Order Consistency (IOC)							RSM Conditional IOC		
	Available Case Means				Rating Scale Model					
Schools	Set 1	Set 2	Set 3	Set 4	Set 2	Set 3	Set 4	Set 2	Set 3	Set 4
1	0.56504	0.57350	0.57015	0.56806	0.57342	0.57398	0.57359	0.50150	0.51881	0.53160
2	0.68484	0.68484	0.68484	0.68219	0.68415	0.68519	0.68183	0.49065	0.54000	0.50000
3	0.64327	0.64119	0.64119	0.64287	0.64420	0.64420	0.64499	0.50972	0.50972	0.51186
4	0.64423	0.64577	0.64557	0.64469	0.64658	0.64658	0.64469	0.50598	0.50598	0.50000
5	0.64499	0.64937	0.64937	0.64104	0.64937	0.64937	0.64528	0.50000	0.50000	0.58182
6	0.61168	0.61346	0.61346	0.61005	0.61257	0.61123	0.60802	0.49537	0.49084	0.49296
7	0.57624	0.57624	0.57327	0.57615	0.58564	0.58564	0.58818	0.56129	0.56129	0.56579
8	0.65689	0.65758	0.65758	0.65518	0.65805	0.65805	0.65448	0.50893	0.50893	0.49580
9	0.61443	0.61443	0.61443	0.61736	0.61931	0.61931	0.62124	0.54054	0.54054	0.52139
10	0.60327	0.61415	0.61415	0.60918	0.62014	0.62014	0.61586	0.55729	0.55729	0.56316
Avg.	0.62449	0.62705	0.62640	0.62468	0.62934	0.62937	0.62782	0.51713	0.52334	0.52644

Table 6. Reliability of Item Measures

	Available Case Means				Rating Scale Model		
Schools	Set 1	Set 2	Set 3	Set 4	Set 2	Set 3	Set 4
1	0.88	0.92	0.92	0.95	0.92	0.94	0.93
2	0.98	0.99	0.99	0.99	0.98	0.98	0.98
3	0.95	0.94	0.94	0.95	0.94	0.94	0.94
4	0.98	0.98	0.98	0.98	0.95	0.96	0.96
5	0.98	0.98	0.98	0.98	0.98	0.98	0.98
6	0.88	0.91	0.91	0.91	0.91	0.92	0.91
7	0.85	0.76	0.81	0.73	0.84	0.89	0.88
8	0.96	0.98	0.98	0.98	0.96	0.97	0.97
9	0.92	0.94	0.94	0.94	0.93	0.93	0.93
10	0.97	0.99	0.98	0.98	0.97	0.97	0.97
Avg.	0.935	0.939	0.943	0.939	0.938	0.948	0.945

Table 7. Reliability of Person Measures

	Rating Scale Model		
Schools	Set 2	Set 3	Set 4
1	0.68	0.64	0.63
2	0.79	0.78	0.77
3	0.76	0.74	0.74
4	0.71	0.68	0.67
5	0.77	0.74	0.74
6	0.64	0.62	0.60
7	0.67	0.60	0.59
8	0.71	0.68	0.68
9	0.74	0.72	0.72
10	0.70	0.68	0.67
Avg.	0.72	0.69	0.68

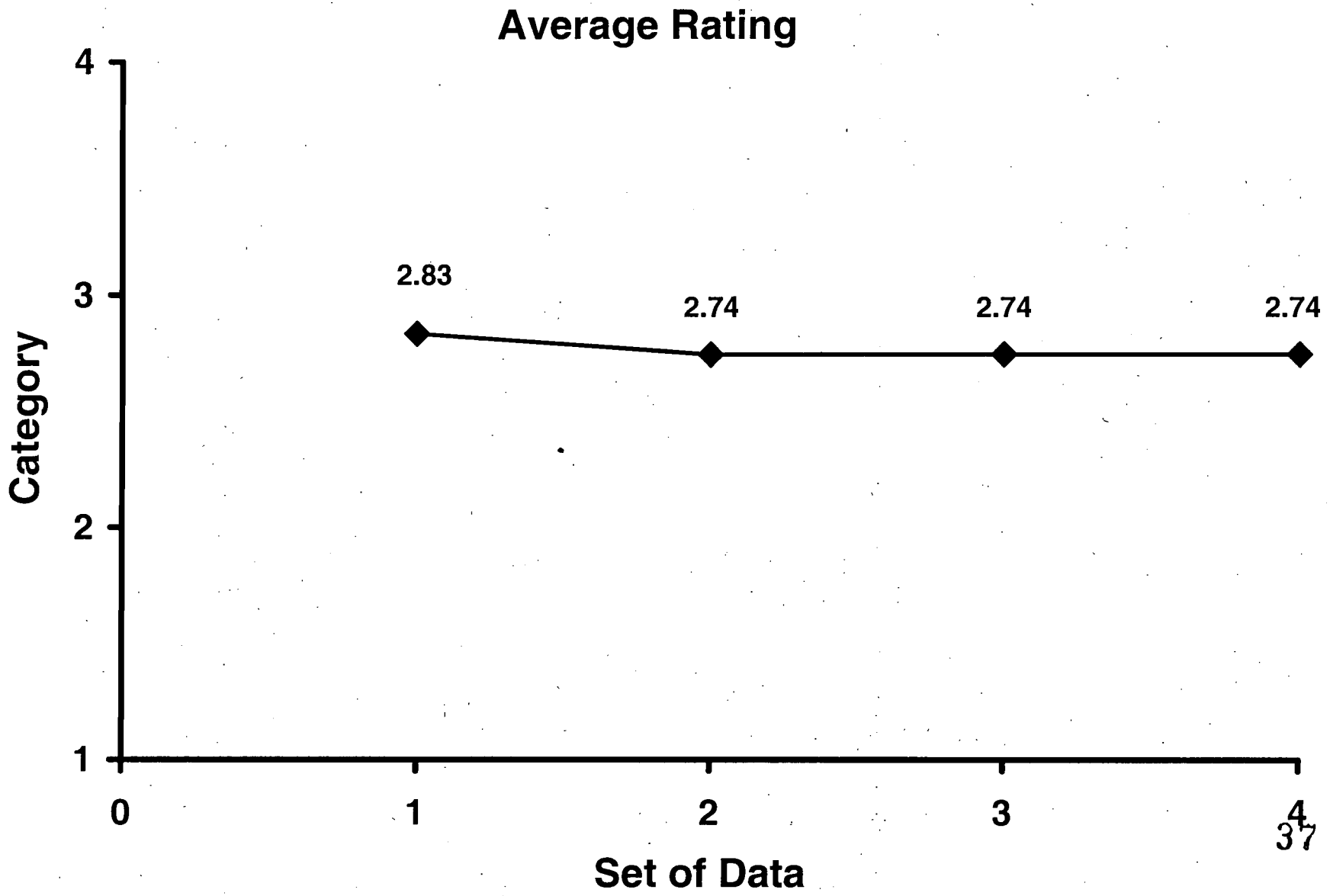
Selected additional tables and figures from the presentation at AERA

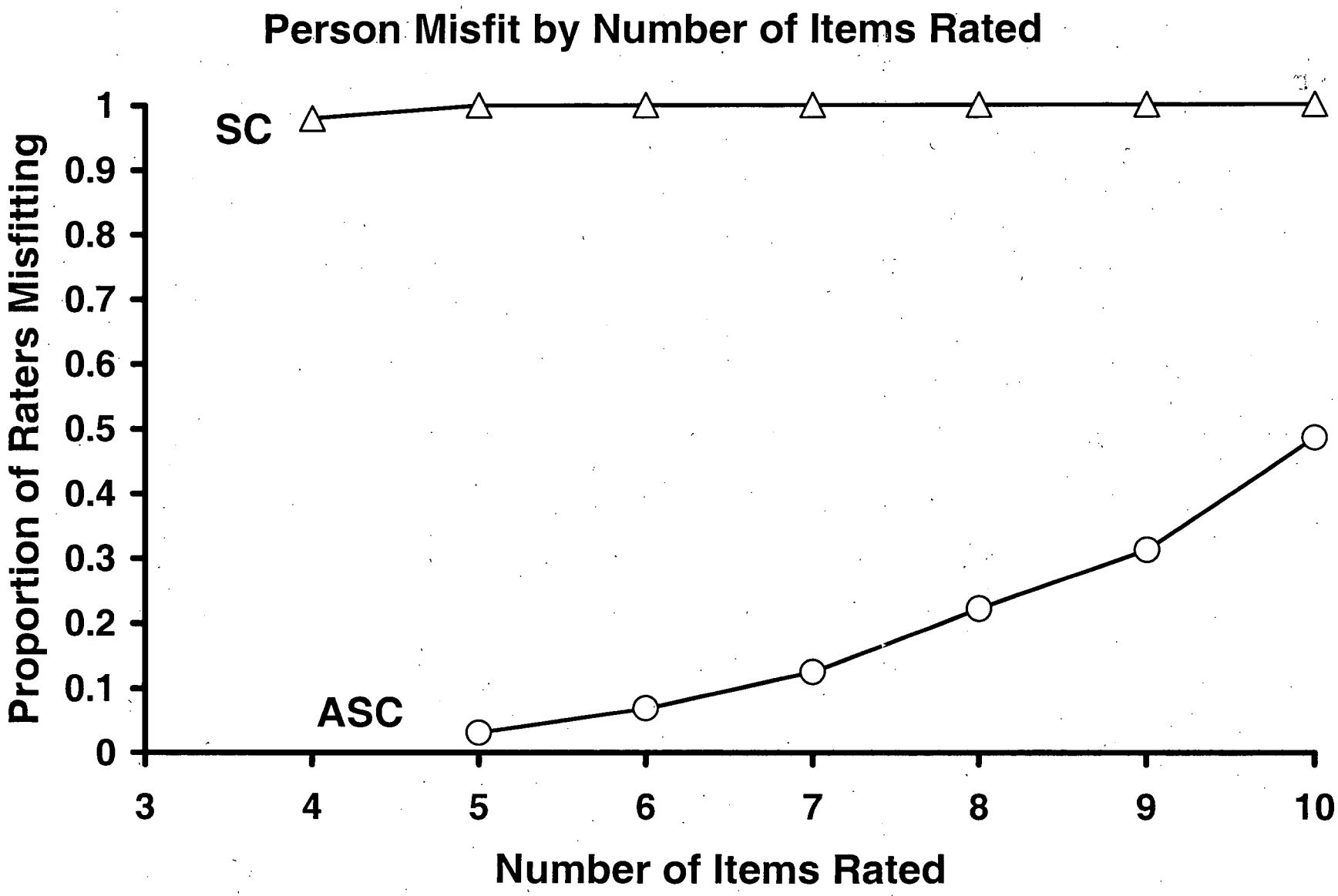
Single Category (SC) Response Sets by Category

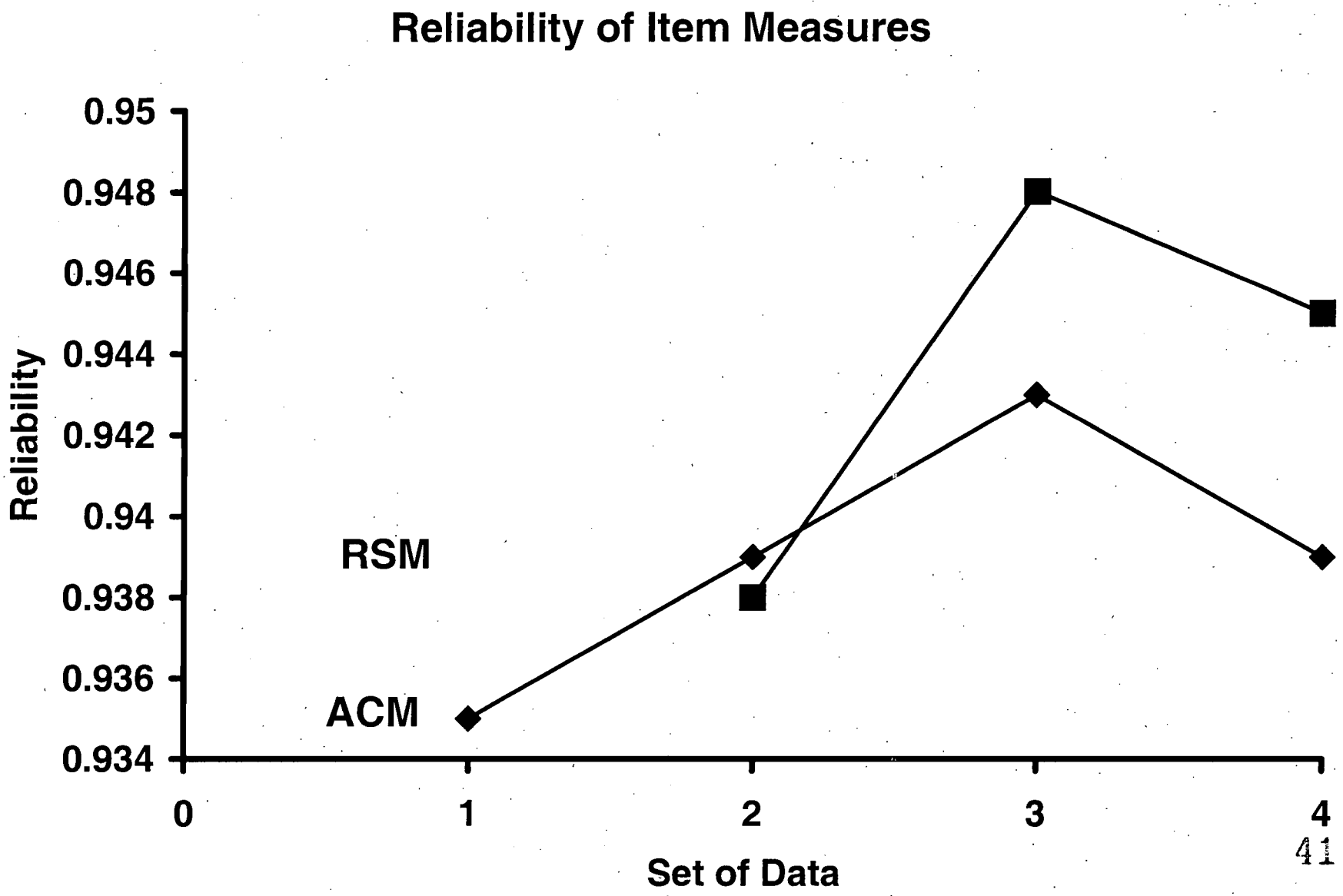
Number of Items Rated	Category 1	2	3	4	All N	All %	Total Raters by Number of Items Rated
10	13	63	135	127	338	37%	924
9	1	3	13	24	41	11%	386
8	0	2	15	20	37	9%	397
7	2	6	14	22	44	9%	501
6	0	8	26	41	75	11%	683
5	4	9	52	42	107	16%	681
4	4	23	67	40	134	22%	622
3	8	32	79	42	161	32%	503
2	12	32	87	43	174	50%	345
1	9	68	92	43	212	100%	212
Total	53	246	580	444	1323	25%	5254

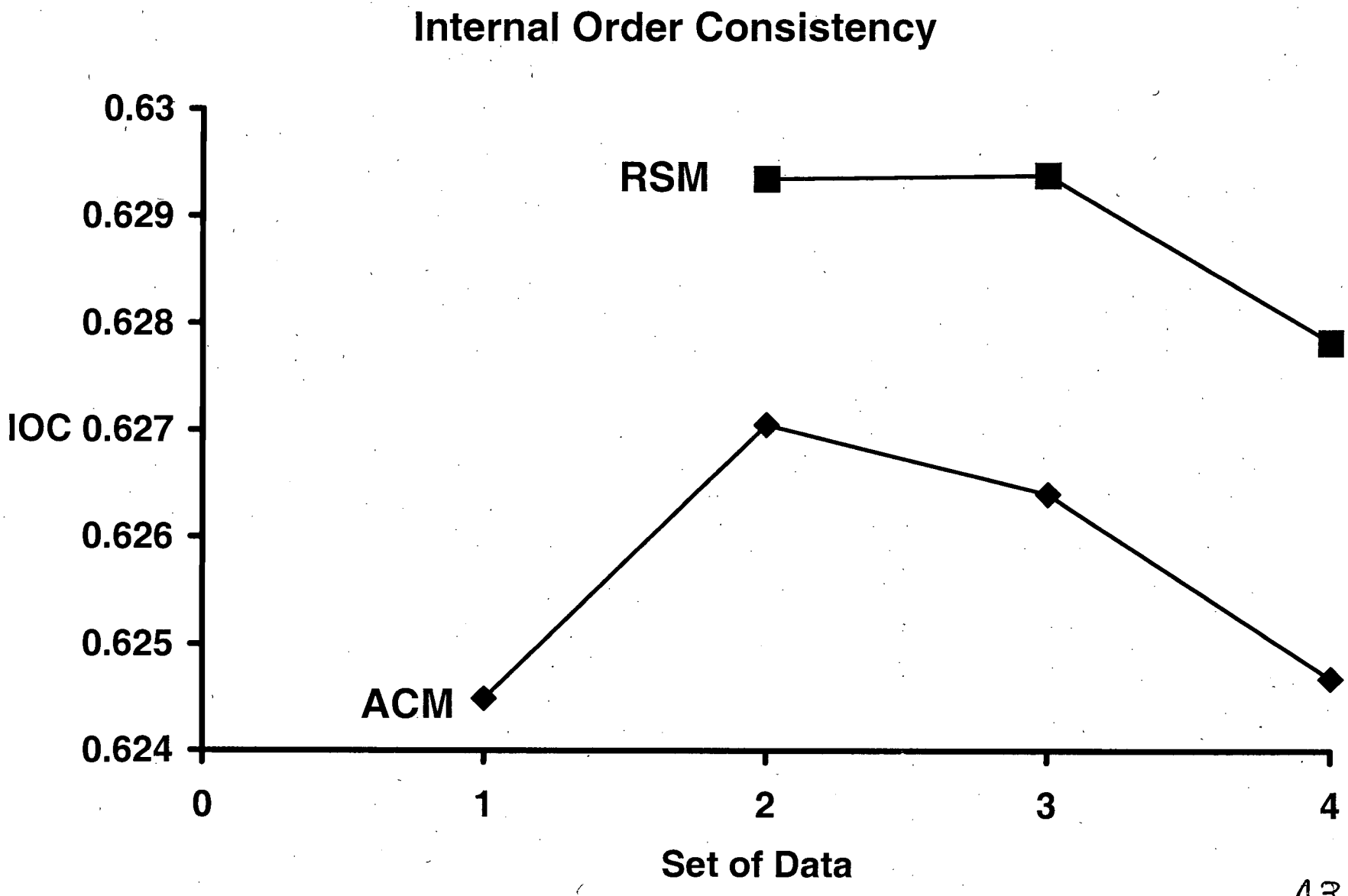
Almost Single Category (ASC) Response Sets by Category

Number of Items Rated	Category 1	2	3	4	All N	All %	Total Raters by Number of Items Rated
10	4	19	29	20	72	8%	924
9	4	5	13	10	32	8%	386
8	6	7	17	15	45	11%	397
7	6	3	24	31	64	13%	501
6	3	22	62	57	144	21%	683
5	12	33	75	43	163	24%	681
4	10	54	114	54	232	37%	622
3	27	84	111	56	278	55%	503
2	--	--	--	--	--	--	345
1	--	--	--	--	--	--	212
Total	72	227	445	286	1030	20	5254


Average Rating
4
3
2
1
Category
2.83
2.74
2.74
2.74
0
1
2
3
4
Set of Data


Person Misfit by Number of Items Rated
Proportion of Raters Misfitting
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
SC
ASC
3
4
5
6
7
8
9
10
Number of Items Rated


Reliability of Item Measures
0.95
0.948
0.946
0.944
0.942
0.94
0.938
0.936
0.934
Reliability
RSM
ACM
0
1
2
3
4
Set of Data


Internal Order Consistency
0.63
0.629
0.628
IOC 0.627
0.626
0.625
0.624
RSM
ACM
0
1
2
3
4
Set of Data

U.S. Department of Education
Office of Educational Research and Improvement (OERI)
National Library of Education (NLE)
Educational Resources Information Center (ERIC)

REPRODUCTION RELEASE

TM032837

(Specific Document)

I. DOCUMENT IDENTIFICATION:

Title: Identifying undifferentiating response sets and assessing their effects on the measurement of items	
Author(s): E. Matthew Schulz and Anji Sun	
Corporate Source: ACT, Inc	Publication Date: April 2001

II. REPRODUCTION RELEASE:

In order to disseminate as widely as possible timely and significant materials of interest to the educational community, documents announced in the monthly abstract journal of the ERIC system, *Resources in Education* (RIE), are usually made available to users in microfiche, reproduced paper copy, and electronic media, and sold through the ERIC Document Reproduction Service (EDRS). Credit is given to the source of each document, and, if reproduction release is granted, one of the following notices is affixed to the document.

If permission is granted to reproduce and disseminate the identified document, please CHECK ONE of the following three options and sign at the bottom of the page.

The sample sticker shown below will be affixed to all Level 1 documents

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY

Sample

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

1

Level 1

☒

Check here for Level 1 release, permitting reproduction and dissemination in microfiche or other ERIC archival media (e.g., electronic) *and* paper copy.

The sample sticker shown below will be affixed to all Level 2A documents

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL IN MICROFICHE, AND IN ELECTRONIC MEDIA FOR ERIC COLLECTION SUBSCRIBERS ONLY, HAS BEEN GRANTED BY

Sample

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

2A

Level 2A

☐

Check here for Level 2A release, permitting reproduction and dissemination in microfiche and in electronic media for ERIC archival collection subscribers only

The sample sticker shown below will be affixed to all Level 2B documents

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL IN MICROFICHE ONLY HAS BEEN GRANTED BY

Sample

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

2B

Level 2B

☐

Check here for Level 2B release, permitting reproduction and dissemination in microfiche only

Documents will be processed as indicated provided reproduction quality permits.
If permission to reproduce is granted, but no box is checked, documents will be processed at Level 1.

I hereby grant to the Educational Resources Information Center (ERIC) nonexclusive permission to reproduce and disseminate this document as indicated above. Reproduction from the ERIC microfiche or electronic media by persons other than ERIC employees and its system contractors requires permission from the copyright holder. Exception is made for non-profit reproduction by libraries and other service agencies to satisfy information needs of educators in response to discrete inquiries.

Sign here, → please

Signature: E. M Schulz	Printed Name/Position/Title: Senior Psychometric Statistician
Organization/Address: P.O. Box 168 Iowa City, IA 52243	Telephone: (319) 337-1468 FAX:
	E-Mail Address: schulz@act.org Date: 5/3/01

(over)