DOCUMENT RESUME

ED 388 678 TM 023 687

AUTHOR Bode, Rita K.

TITLE Using Rasch To Create Measures from Survey Data (or Making a Silk Purse out of a Sow's Ear).

PUB DATE Apr 95

NOTE 16p.; Paper presented at the Annual Meeting of the Educational Research Association (San Francisco, CA, April 18-22, 1995).

PUB TYPE Reports - Evaluative/Feasibility (142) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC01 Plus Postage.

DESCRIPTORS *Ability Grouping; Classification; Educational Practices; *Item Response Theory; *Mathematics Teachers; Measurement Techniques; *Surveys; *Test Construction; Test Items

IDENTIFIERS BIGSTEPS Computer Program; *Composite Scores; Dimensionality (Tests); Missing Data; *Rasch Model; Second International Mathematics Study

ABSTRACT

This study describes the creation of measures of teachers' use of ability grouping in instruction using Rasch analysis. The dimensionality of the proposed construct was also investigated. Results of the Rasch analysis are compared to the results using composites to illustrate how the description of a construct can vary depending on the method used to create its measure. The sample consisted of 299 eighth-grade mathematics teachers who participated in the 1981-82 Second International Mathematics Study. Teachers responded to a Teacher's Questionnaire and a General Classroom Process Questionnaire. Items selected from these surveys were those with relevance for the organization of students for instruction. Raw data was categorized and calibrated using the BIGSTEPS computer program. The same data were used in the creation of composite scores. In terms of reliability, approximately comparable results were found for the Rasch measures and the traditional composites. Rasch also made it possible to see the hierarchy of practices that formed the continuum on which estimates of teachers' position were based. In addition, the ability to deal with missing data made Rasch more useful. An appendix discusses some aspects of grouping arrangements. (Contains six tables, four figures, and three references.) (SLD)

Using Rasch to Create Measures from Survey Data (or Making a Silk Purse out of a Sow's Ear)

Rita K. Bode

University of Illinois at Chicago

Paper presented at the Annual Meeting of the
American Educational Research Association
San Francisco CA, April 1995

Using Rasch to Create Measures from Survey Data
(or Making a Silk Purse out of a Sow's Ear)

Discussions of the creation of measures of constructs (Wright & Stone, 1979) usually deal with situations in which instruments are developed from scratch and can include any or all potential indicators of a construct. However, when doing secondary data analysis using survey data, researchers are restricted to items included in the original survey. While there are benefits to using survey data in terms of sample representativeness and size, one disadvantage is the inability to control the content and structure of the items included. Even though the survey may contair various indicators of the construct, items that deal with specific aspects of the research question may not be included. Also, the structure of the scales used in these items may not be compatible.

Traditionally, if the same response scale is used for all items, values on this scale can be summed or averaged across items to create a composite measure. If the response scale varies across items, it can be transformed into z scores prior to the creation of the composite measure. The reliability of the composite is used to determine the quality of the scale. In statistical packages such as SPSS, Cronbach's alpha is used to determine the internal consistency of the set of items. In addition to the overall estimate of the reliability of the composite, the composite is recalculated with each individual item deleted and the reliability of these sets of items is provided. This information can be used to identify items that decrease the internal consistency of the set of items. In turn, these items can be deleted from the set to increase the reliability of the composite. In some instances, LISREL can be used to create a more reliable measure of the composite. An approach to creating measures from survey data that is seldom described is the use of Rasch analysis.

The construct under investigation in this study is teachers' use of ability grouping for instruction. In previous research in this area, the practice of ability grouping has been defined dichotomously: either it is used or not used. However, grouping arrangements may differ from class to class in the way in which student groups are organized. Also once arranged in groups, students may receive different instruction depending on the group into which they are placed. Since, according to Gamoran (1987), grouping does not produce achievement--instruction does, accounting for variation in both aspects of practice is important. The use/nonuse dichotomy cannot not represent a clear description of teachers' practice concerning ability grouping.

In general, according to Gamoran & Behrends (1987), the use of a continuous variable probably reflects a construct more accurately than the use of a dichotomy. A continuous scale along which teachers are positioned relative to various organizational arrangements and instructional practices would seem to be advantageous in that it could encompass various types of practice within the same scale. At one end of the continuum would be teachers who incorporate many different types of practice that are characteristic of ability grouping and at the other would be teachers who do not incorporate these practices into their instruction at all. The indicators used to describe this construct would also form a continuum. At one end would be practices which many teachers include in their instruction and at the other end would be practices which few teachers--only those who use grouping by ability to a substantial extent--would include in their instruction.

This study describes the creation of measures of teachers' use of ability grouping in instruction using Rasch analysis. The dimensionality of the proposed construct is also investigated. Not only is a continuous measure a more accurate indicator of teachers' practice than the use of a dichotomy, but the use of Rasch analysis is expected to provide a fuller description of the construct than using traditional composite scores. The results of the Rasch analysis are compared to the results using composites to illustrate how the description of a construct can vary depending on the method used to create its measure.

METHODOLOGY

Sample

The sample consists of 299 eighth-grade mathematics teachers who participated in the 1981-82 Second International Mathematics Study (SIMS). The vast majority of these teachers taught mathematics classes described as "typical" but the sample also includes data for teachers who taught remedial, enriched, and algebra classes.

Instrument

The SIMS teachers responded to a Teacher's Questionnaire and a General Classroom Process Questionnaire. The items on these surveys cover demographic characteristics of the teachers and the classrooms, attitudes of the teachers, and descriptions of classroom process. The items selected from these surveys for this study are those which have relevance for the organization of students for instruction and some aspects of the instruction provided. Some of the items refer to teachers' beliefs about the importance of various practices while others refer to actual organizational and instructional practices. A description of the selected items and their response scales is presented in Appendix A.

Analysis

Prior to Rasch calibration, categorization of the some of the raw data is needed to abstract the meaning of the responses. One set of items requests estimates of the amount or percentage of time spent in various activities or arrangements. The range of responses to these items is from 0 to 100 percent. Since there are not 101 distinct levels of time spent in an activity or arrangement, categorization into meaningful categories is needed. Frequency distributions of the responses to individual items are used to identify different levels of time spent. Cutoff points are used to transform the time estimates into categories in which a "0" indicates no time in the grouping arrangement, "1" indicates a minimal amount of time (1-33%) in the grouping arrangement, "3" indicates a moderate amount of time (34-66%) in the grouping arrangement, and "4" indicates a predominant amount of time (67-100%) in the grouping arrangement.

Other items require no such categorization. Some of these items are dichotomous: a practice was either used or not used (e.g., is pacing varied, are students grouped by ability, etc.). Other items are polychotomous: beliefs about the importance of various practices are rated on a scale from "0" (not at all important) to "4" (of utmost importance). In terms of grouping arrangements, since the opposite of ability grouping may not be whole class instruction but other grouping arrangements such as mixed ability grouping, new items are created to identify the practice of mixed ability grouping and no grouping at all.

Once categorized, these items are calibrated using BIGSTEPS. Since the items selected tap into various aspects of teacher practice, it is possible that the selected items measure more than one construct. In order to determine whether grouping and tailoring practice represents a single or multiple constructs, BIGSTEPS is run on different sets of items. In the initial BIGSTEPS run, all items are included into a single measure. In subsequent runs, the grouping and tailoring items are split and separate calibrations are obtained for each set of items. In a final set of runs, the items are further split into those regarding the type of grouping arrangement used, the time spent in various grouping arrangements, tailoring beliefs and tailoring practices. Results of these calibrations are compared to determine the optimal number and composition of measures.

These same data are used in the creation of composite scores. In the first analysis, all items are combined into a single composite and in subsequent runs, the grouping and tailoring items are split and separate composites are created. Using the RELIABILITY feature of SPSS, the quality of the composites is determined and the results are used to determine whether grouping and tailoring represent a single or dual constructs.

RESULTS

The results from the initial calibration are presented in Table 1 and shown in Figure 1 (on page 9). Table 1 presents item and person summary statistics for all items combined and Figure 1 shows the item and person map for this calibration.

Table 1

Summary Statistics for Grouping and Tailoring Items Combined

```
   SUMMARY OF    233 MEASURED (NON-EXTREME) PERSONS
 ------------------------------------------------------------------------------
|           RAW                                          INFIT        OUTFIT   |
|           SCORE      COUNT      MEASURE    ERROR     MNSQ    STD   MNSQ   STD |
|------------------------------------------------------------------------------|
| MEAN      24.1       19.1         -.90      .50      1.03    -.1    .95   -.2 |
| S.D.       6.2        1.9         1.36      .09       .47    1.2    .66   1.1 |
|------------------------------------------------------------------------------|
| RMSE    .51   ADJ.S.D.    1.26    SEPARATION  2.49   PERSON RELIABILITY   .86 |
 ------------------------------------------------------------------------------
    LACKING RESPONSES:      2 PERSONS

   SUMMARY OF     20 MEASURED (NON-EXTREME) ITEMS
 ------------------------------------------------------------------------------
|           RAW                                          INFIT        OUTFIT   |
|           SCORE      COUNT      MEASURE    ERROR     MNSQ    STD   MNSQ   STD |
|------------------------------------------------------------------------------|
| MEAN     280.5      222.3          .00      .16       .97    -.2    .93   -.1 |
| S.D.     279.1       13.2         1.28      .05       .19    1.9    .27   1.8 |
|------------------------------------------------------------------------------|
| RMSE    .17   ADJ.S.D.    1.27    SEPARATION  7.55    ITEM  RELIABILITY   .98 |
 ------------------------------------------------------------------------------
```

When combined, how well do grouping and tailoring items function to describe a continuum? After deleting four items because of misfit--individualized instruction time, student time in seatwork or blackboard work, the use of mixed-ability grouping, and the practice of varying discussion questions in class--the results of the initial calibration appear promising. Person and item separation are good while the item misfit is only slightly high.

What, however, is the continuum defined by this combination of items? The person and item map for the items combined proves difficult to interpret because of the inclusion of so many different types of items. The items appear to cluster into two groups which can be thought of as more common and less common practices. Among the more common practices are no use of grouping at all and beliefs about tailoring. Among the less common practices are the use of grouping by ability and various practices of tailoring instruction and assignments. Those items dealing with time spent in grouping arrangements and various practices of tailoring assignments lean toward less common use. The interweaving of grouping and tailoring items suggests that separation of items into at least two sets might improve the interpretability of the construct.

The results from the separate calibrations are presented in Tables 2 and 3 and shown in Figures 2 and 3 (on pages 10 and 11). Table 2 presents item and person summary statistics for the grouping items and Figure 2 shows the item and person map for this calibration. Table 3 presents summary statistics for the tailoring items and Figure 3 shows the map for this calibration.

Table 2

Summary Statistics for Grouping Items

```
    SUMMARY OF    227 MEASURED (NON-EXTREME) PERSONS
 -------------------------------------------------------------------------------
|          RAW                                          INFIT        OUTFIT      |
|          SCORE       COUNT       MEASURE    ERROR    MNSQ    STD   MNSQ    STD |
|-------------------------------------------------------------------------------|
| MEAN       7.5         7.8        -1.06      .82      .99    -.3    .90    -.3 |
| S.D.       2.9          .7         1.80      .13      .76    1.1    .81     .8 |
|-------------------------------------------------------------------------------|
| RMSE    .83    ADJ.S.D.     1.59    SEPARATION   1.92    PERSON RELIABILITY  .79 |
 -------------------------------------------------------------------------------
 MINIMUM EXTREME SCORE:      6 PERSONS
     LACKING RESPONSES:      2 PERSONS

    SUMMARY OF     8 MEASURED (NON-EXTREME) ITEMS
 -------------------------------------------------------------------------------
|          RAW                                          INFIT        OUTFIT      |
|          SCORE       COUNT       MEASURE    ERROR    MNSQ    STD   MNSQ    STD |
|-------------------------------------------------------------------------------|
| MEAN     213.6       222.3          .00'     .16      .99    -.1    .89    -.5 |
| S.D.     179.0         4.8         1.32      .04      .19    2.0    .26    1.6 |
|-------------------------------------------------------------------------------|
| RMSE    .17    ADJ.S.D.     1.31    SEPARATION   7.95     ITEM  RELIABILITY  .98 |
 -------------------------------------------------------------------------------
```

Table 3

Summary Statistics for Tailoring Items

```
    SUMMARY OF    228 MEASURED (NON-EXTREME) PERSONS
 -------------------------------------------------------------------------------
|          RAW                                          INFIT        OUTFIT      |
|          SCORE       COUNT       MEASURE    ERROR    MNSQ    STD   MNSQ    STD |
|-------------------------------------------------------------------------------|
| MEAN      13.8        11.4         -.96      .70     1.02    -.1   1.02    -.2 |
| S.D.       3.6          .9         1.60      .11      .60    1.2   1.38    1.0 |
|-------------------------------------------------------------------------------|
| RMSE    .71    ADJ.S.D.     1.43    SEPARATION   2.02    PERSON RELIABILITY  .80 |
 -------------------------------------------------------------------------------
 MAXIMUM EXTREME SCORE:      3 PERSONS
     LACKING RESPONSES:      4 PERSONS

    SUMMARY OF    12 MEASURED (NON-EXTREME) ITEMS
 -------------------------------------------------------------------------------
|          RAW                                          INFIT        OUTFIT      |
|          SCORE       COUNT       MEASURE    ERROR    MNSQ    STD   MNSQ    STD |
|-------------------------------------------------------------------------------|
| MEAN     261.6       216.0          .00      .18      .96    -.3    .99     .0 |
| S.D.     277.2        14.8         1.51      .05      .19    1.8    .40    1.7 |
|-------------------------------------------------------------------------------|
| RMSE    .19    ADJ.S.D.     1.49    SEPARATION   7.98     ITEM  RELIABILITY  .98 |
 -------------------------------------------------------------------------------
```

Does sèparating grouping and tailoring items produce an improvement in person and item measurement? Two criteria are used to determine whether the separate calibrations produce an improvement: person separation and model fit. Comparing the results of these calibrations with those for all items combined shows that, for grouping and tailoring items separately, person separation decreases slightly (grouping: from .83 to .79; tailoring: from .83 to .80) while item misfit remains essentially the same. Using these criteria, it appears that nothing is gained by calibrating the items separately. Grouping and tailoring practice are but two aspects of the same construct.

Another criterion for determining how useful the results of a calibration are is the interpretation of the item map. The item maps for the separate calibrations prove easier to interpret. The relative positions of the grouping item remain essentially the same as before but, by restricting the content to just those practices dealing with grouping, the interpretation of the continuum becomes clearer. The least common practice involves grouping by ability and grouping the least able students together and having students spend a predominant proportion of time in small group work. More common is the practice of grouping the most able students together and spending a predominant proportion of time in small group instruction and most common is not grouping at all.

To fit the Rasch model, the scales on several items need to be reversed. These items are: whole class instruction time, student time in whole class work, and no grouping at all. The reversal of the scales for time spent in whole class instruction and student time in whole class work makes the interpretation trickier. In terms of time spent in different types of grouping arrangements, having students spend a predominant proportion of time in small group work is less common than having them spend a minimal proportion of time in whole class work (e.g., listening to whole class lectures) but there is little differentiation between spending a predominant proportion of time in small group instruction and spending a minimal proportion of time in whole class instruction. These results suggest that the use of a combination of grouping arrangements falls somewhere between using and not using small group instruction.

The relative position of the tailoring items also remained essentially the same as when the two item types were combined but the isolation of item content dealing with tailoring makes for easier interpretation. The tailoring item map shows a break between the belief and practice. In terms of beliefs, teachers are more likely to believe that more able students should be given harder tasks than that less able students should be given easier tasks. In terms of practice, teachers are more likely to: 1) tailor assignments than instruction, 2) vary assignment due dates than the assignments themselves, 3) vary pacing in instruction rather than content, and 4) assign harder exercises than harder topics. They are also less likely to vary assignments frequently or assign more exercises to some students.

Further separation of items into the different aspects of grouping and tailoring was explored. For the grouping items, this breakdown was into grouping type used and time spent in various grouping arrangements and for tailoring, it was into tailoring beliefs and practices. The results of this investigation were mixed in terms of whether there was any improvement in person separation and model fit. Therefore, exploration into the further separation of items was discontinued.

Separation of the items into grouping and tailoring produces more interpretable continua while the combined calibrations produce greater separation and essentially the same model fit. How then does one decide which calibrations to use to define grouping/tailoring construct(s)? One way is to look at the relationship between person measures from the separate calibrations to see if teachers who are high on one measure are high on the other. If this relationship is strong, one can assume that the two sets of items are measuring the same construct; if the relationship is weak, one can assume that they are measuring two separate constructs.

The correlation between the teachers' measures from the separate calibrations shows only a moderate positive relationship (r = .551). Figure 4 (on page 12), a plot of the calibrations for the two measures, shows that teachers who practice ability grouping do not necessarily tailor instruction to the same extent. The measures on these two constructs for many teachers are strongly related: those who group for ability also tailor instruction and those who don't use grouping at all don't even believe in tailoring. But some teachers practice ability grouping but not instructional tailoring and others practice instructional tailoring but not ability grouping. Since it is instruction and not grouping per se that produces learning, perhaps these results explain why ability grouping doesn't always have an effect on subsequent student achievement.

Because this level of relationship is moderate, the information on whether these items represent one or two constructs is not conclusive. However, due to the improvement in the interpretability of the continua resulting from the separate calibration of grouping and tailoring items and the fact that the relationship between the two measures is only moderate, it appears that treating grouping and tailoring as separate constructs is preferable.

Finally, traditional composites were created for grouping and tailoring items separately and combined. One major stumbling block in using traditional composites is the requirement of complete data for all subjects. Due to this restriction, 100 cases were dropped from the analysis--almost half of the sample. Because of peculiarities in these data, this restriction also resulted in an inability to obtain alpha coefficients. It happened that every one of the teachers who responded positively to the tailoring of instruction or assignment items also had missing responses; therefore, when the cases with missing data were eliminated, all the remaining responses to these items were zero. With zero variance for these items, it was not possible to calculate correlation coefficients and subsequent alpha coefficients.

With manipulation of the data, however, it is possible to obtain the reliability data. Missing responses are replaced with zero responses with the assumption that teachers who practice an aspect of tailoring would have responded "yes" and those who do not practice that aspect could have responded "no" or left the item blank. An initial run was made using all items. A subsequent run was made deleting those items identified as correlating poorly with the composite; that is, those items whose deletion would result in an increase in the alpha coefficient. The results of this subsequent analysis are presented in Tables 4 to 6.

Table 4

Reliability Analysis for Grouping/Tailoring Composite

```
# OF CASES =        212.0
                                                 # OF
STATISTICS FOR        MEAN    VARIANCE    STD DEV  VARIABLES
    SCALE            -.1430   108.9114    10.4361       20

ITEM MEANS            MEAN     MINIMUM    MAXIMUM      RANGE     MAX/MIN    VARIANCE
                     -.0072    -.0489      .0231       .0719     -.4719       .0003

ITEM VARIANCES        MEAN     MINIMUM    MAXIMUM      RANGE     MAX/MIN    VARIANCE
                      .9799     .8670     1.0470       .1801     1.2077       .0014

INTER-ITEM
COVARIANCES           MEAN     MINIMUM    MAXIMUM      RANGE     MAX/MIN    VARIANCE
                      .2350    -.1769      .6998       .8767    -3.9549       .0152

INTER-ITEM
CORRELATIONS          MEAN     MINIMUM    MAXIMUM      RANGE     MAX/MIN    VARIANCE
                      .2399    -.1823      .7021       .8844    -3.8515       .0158

ITEM-TOTAL
STATISTICS     SCALE          SCALE      CORRECTED
               MEAN         VARIANCE       ITEM-        SQUARED         ALPHA
              IF ITEM        IF ITEM       TOTAL        MULTIPLE       IF ITEM
              DELETED        DELETED    CORRELATION   CORRELATION      DELETED

ZSGINST        -.1555        96.0677       .5995         .7102          .8514
ZWCINST        -.1369        96.6354       .5794         .5287          .8522
ZGRPWORK       -.1480        97.3621       .5373         .4609          .8538
ZCLASWRK       -.1444        98.6635       .4614         .4990          .8568
ZABILGRP       -.1295        98.5586       .4785         .3134          .8561
ZMOSTABL       -.1421       100.7354       .3575         .3188          .8607
ZLESTABL       -.1476       100.6280       .3612         .1970          .8606
ZNOGRPG        -.1593        99.9278       .4044         .6144          .8589
ZPACING        -.1661       100.1200       .3782         .4478          .8600
ZCONTENT       -.0942       100.5492       .4014         .4322          .8589
ZDUEDATE       -.1368       105.1517       .1361         .3361          .8687
ZASSIGN        -.1145        96.6253       .5900         .6840          .8519
ZMOREXER       -.1076       100.4475       .3922         .3160          .8593
ZHARDEX        -.1248        95.6896       .6314         .6427          .8502
ZHARDTOP       -.1242        97.4561       .5396         .4547          .8538
ZSIMPLE        -.1143        99.6012       .4162         .3132          .8585
ZHARDER        -.1497        99.3294       .4374         .3759          .8577
ZVHARDEX       -.1521       100.4533       .3820         .3872          .8597
ZTAILASG       -.1259        97.9339       .5071         .3911          .8550
ZFREQ          -.1440        98.0673       .5030         .3121          .8552

RELIABILITY COEFFICIENTS     20 ITEMS
ALPHA =    .8632            STANDARDIZED ITEM ALPHA =    .8633
```

The reliability coefficients for these composites are comparable to the person separation reliability obtained from the Rasch analyses (.86 for grouping and tailoring combined, .79 for grouping, and .80 for tailoring). The same items that are poorly correlated with the composites were identified as misfitting to the Rasch model. But would we have drawn the same conclusion regarding whether to combine the grouping and tailoring items or separate them? The results of the reliability analysis would indicate that the combined composite was preferable since the reliability coefficient is highest. Whether one would have taken the extra step to investigate the relationship between grouping and tailoring composites in questionable. Most likely, with such a high reliability coefficient, one would have just used the combined composite without further investigation.

If one did look at the relationship between the two composites, one would have found that the relationship was moderate (.614). This coefficient is slightly higher than the correlation between the two measures which indicate: that the use of composites slightly exaggerates the relationship between these two constructs. The plot of the grouping and tailoring (not shown) appears to be relatively similar to the plot of the Rasch measures and, as such, should lead to a similar decision in terms of whether grouping and tailoring items should be combined or separated.

Table 5

Reliability Analysis for Grouping Composite

```
# OF CASES =        212.0
                                                      # OF
STATISTICS FOR         MEAN   VARIANCE    STD DEV  VARIABLES
     SCALE            .0193    24.7244     4.9724        8

ITEM MEANS             MEAN    MINIMUM    MAXIMUM      RANGE    MAX/MIN   VARIANCE
                      .0024    -.0135      .0163      .0298    -1.2100      .0001

ITEM VARIANCES         MEAN    MINIMUM    MAXIMUM      RANGE    MAX/MIN   VARIANCE
                      .9957     .9746     1.0149      .0403     1.0413      .0002

INTER-ITEM
COVARIANCES            MEAN    MINIMUM    MAXIMUM      RANGE    MAX/MIN   VARIANCE
                      .2993     .1145      .6998      .5852     6.1101      .0202

INTER-ITEM
CORRELATIONS           MEAN    MINIMUM    MAXIMUM      RANGE    MAX/MIN   VARIANCE
                      .3006     .1148      .7021      .5873     6.1141      .0202

ITEM-TOTAL
STATISTICS     SCALE           SCALE      CORRECTED
               MEAN          VARIANCE      ITEM-         SQUARED           ALPHA
              IF ITEM        IF ITEM       TOTAL         MULTIPLE         IF ITEM
              DELETED        DELETED    CORRELATION    CORRELATION        DELETED

ZSGINST        .0068         18.1482       .6485          .6445            .7198
ZWCINST        .0254         19.0405       .5434          .4367            .7388
ZGRPWORK       .0143         19.0319       .5411          .4050            .7392
ZCLASWRK       .0178         19.3914       .4867          .4581            .7485
ZABILGRP       .0327         19.9952       .4253          .1978            .7588
ZMOSTABL       .0201         20.5581       .3492          .1941            .7714
ZLESTABL       .0147         20.8506       .3125          .1246            .7774
ZNOGRPG        .0029         19.2955       .5115          .5555            .7443

RELIABILITY COEFFICIENTS     8 ITEMS
ALPHA =    .7747            STANDARDIZED ITEM ALPHA =    .7747
```

Table 6

Reliability Analysis for Tailoring Composite

```
# OF CASES =        212.0
                                                      # OF
STATISTICS FOR         MEAN   VARIANCE    STD DEV  VARIABLES
     SCALE           -.1623    42.9575     6.5542       12

ITEM MEANS             MEAN    MINIMUM    MAXIMUM      RANGE    MAX/MIN   VARIANCE
                     -.0135    -.0489      .0231      .0719     -.4719      .0004

ITEM VARIANCES         MEAN    MINIMUM    MAXIMUM      RANGE    MAX/MIN   VARIANCE
                      .9693     .8670     1.0470      .1801     1.2077      .0019

INTER-ITEM
COVARIANCES            MEAN    MINIMUM    MAXIMUM      RANGE    MAX/MIN   VARIANCE
                      .2373    -.1769      .6386      .8155    -3.6092      .0223

INTER-ITEM
CORRELATIONS           MEAN    MINIMUM    MAXIMUM      RANGE    MAX/MIN   VARIANCE
                      .2450    -.1823      .6597      .8420    -3.6186      .0237
ITEM-TOTAL
STATISTICS     SCALE           SCALE      CORRECTED
               MEAN          VARIANCE      ITEM-         SQUARED           ALPHA
              IF ITEM        IF ITEM       TOTAL         MULTIPLE         IF ITEM
              DELETED        DELETED    CORRELATION    CORRELATION        DELETED

ZPACING       -.1853         37.6686       .3377          .3729            .7909
ZCONTENT      -.1134         38.0113       .3553          .3485            .7885
ZDUEDATE      -.1561         41.1211       .0667          .2739            .8152
ZASSIGN       -.1337         35.4640       .5620          .6421            .7688
ZMOREXER      -.1268         37.4543       .3901          .2781            .7854
ZHARDEX       -.1440         34.3824       .6535          .5972            .7593
ZHARDTOP      -.1434         35.7923       .5261          .4260            .7723
ZSIMPLE       -.1335         36.7198       .4320          .2716            .7815
ZHARDER       -.1689         36.0790       .4978          .3530            .7751
ZVHARDEX      -.1713         36.8220       .4355          .3800            .7812
ZTAILASG      -.1451         35.6729       .5296          .3719            .7719
ZFREQ         -.1633         36.0196       .5012          .2781            .7747

RELIABILITY COEFFICIENTS     12 ITEMS
ALPHA =    .7955            STANDARDIZED ITEM ALPHA =    .7956
```

Conclusions

Two issues have been addressed in this study. The first is how to decide whether the set of items one is working with represents a single or multiple constructs and the second is the effect of using Rasch analysis as compared to traditional composites on the decision made. In situations where the number of constructs represented by a set of items is unclear, separate calibration and comparison of the resulting person measures can be informative. If the two measures are measuring the same construct, they should be highly correlated and the plot points should fall along the identity line. If the two measures are measuring different constructs, the relationship between measures should be weaker and the plot points more dispersed. The size of the correlation coefficient and dispersion of the plot points should provide guidance as to the number of constructs involved.

Would the decision as to whether one or two constructs were represented by the set of items be affected by the method used to analyze the data. From the results of this study, perhaps different decisions would have been made. Using traditional composites, the combined set of grouping and tailoring items produced a higher reliability (by virtue of the greater number of items) and most probably would be selected. Using Rasch calibrations, the combined measures for grouping and tailoring measure also produced greater person reliability with essentially the same amount of misfit. However, the difficulty in interpreting the resulting continuum would probably lead one to select the measures from the separate calibrations.

What can one conclude about the use of Rasch measures instead of traditional composites in creating measures from survey data? In terms of reliability, approximately comparable results are found. However, Rasch provides the structure to enable one to look at the composition of the measures which is not available with traditional composites. Using Rasch, one can see the hierarchy of practices that form the continuum upon which estimates of teacher's position are based. Looking at the content of the items on this continuum provides qualitative information upon which to make decisions concerning dimensionality. With composites, all one knows is that the responses to the set of items are internally consistent.

More importantly, especially in this case, the ability to deal with missing data makes Rasch more useful in creating measures of constructs. In the least not being able to handle missing data decreases the size of the sample one is using; at most, it may prevent one from determining the quality of the composite created. In this case, it was possible to replace the missing data with zero scores and be reasonably confident that the meaning of the responses was not changed, but had the missing data been not been in items that were dichotomous, this adjustment would not have been possible.

Researchers don't always have control over the content of surveys used to collect data in their specific area of interest and may need to create measures using whatever data is available. Rather than using a dichotomy to describe the presence or absence of a practice, a continuum along which people vary can be created using various indicators of the practice. Instead of creating a traditional composite from these indicators, this study shows how indicators can be created and used in a Rasch analysis to obtain a useful and meaningful measure that can enhance understanding of the construct under study.

References

Gamoran, A. (1987). Organization, instruction, and the effects of ability grouping: Comment on Slavin's "best-evidence synthesis." *Review of Educational Research*, 57(3), 341-345.

Gamoran, A., & Behrends, M. (1987). The effects of stratification in secondary schools: Synthesis of survey and ethnographic research. *Review of Educational Research*, 57(4), 415 435.

Wright, B.D., & Stone, M.H. (1979). *Best Test Design*. Chicago: MESA Press.

Figure 1

Person and Item Map for Grouping and Tailoring Items Combined

```
         PERSONS        MAP OF ITEMS

      4             +                                                                    <--- LESS COMMON

                   .

      3            . +



Q
                  .# |
                   # |
      2            . +
                   . |
                  .# | C39 SOME ASSIGNED MORE EXERCISES
                   # | C37B CONTENT VARIED
S                 .# | C34 GROUPING BY ABILITY
                   . | C33 LEAST ABLE GROUPED TOGETHER
      1           .# + C37A PACING VARIED                        C41 SOME ASSIGNED HARDER TOPICS
                 ### | T24C STUDENT TIME IN GROUP WORK
                .### | C38A DUE DATES VARIED                     C38B ASSIGNMENTS VARIED
                 ### |
                  .# | C40 SOME ASSIGNED HARDER EXERCISES
             ####### | C28 [REV] WHOLE CLASS INSTRUCTION TIME    C29 SMALL GROUP INSTRUCTION TIME
M     0          ### + T26 [REV] FREQUENCY OF DIFFERENT ASSIGNMENTS
               ##### | C32 MOST ABLE GROUPED TOGETHER            T24B [REV] STUDENT TIME IN CLASS WORK
                 ### |
             ####### |
               ##### |
                  .# |
     -1         .### +
             ####### | C62 IMP: SIMPLER TASKS FOR LESS ABLE
S        .########## |
                     |
           .######## | C97 IMP: ASSIGNMENTS TAILORED TO NEEDS
        .########### |
     -2            . + C91 IMP: VERY HARD TASKS FOR TRULY ABLE
             ####### |
                     |
Q             ###### | C36 [REV] NO GROUPING                     C67 IMP: HARDER TASKS FOR MORE ABLE
                     |
                #### |
     -3              +
                  .# |
                     |
                 ### |

     -4            . +                                                                    <--- MORE COMMON
```

Figure 2

Person and Item Map for Grouping Items

```
     PERSONS       MAP OF ITEMS

  4          . +                                        <--- LESS COMMON
               |
               .
           . . |
  3            +
               |
Q            . |
               |
           #   |
  2            +
               |
          .#   |
               |
S        .##   | C34 GROUPING BY ABILITY
               | C33 LEAST ABLE GROUPED TOGETHER
  1        #   + T24C STUDENT TIME IN GROUP WORK
       .####   |
           .   |
     #######   |
           .   |
               | C28 [REV] WHOLE CLASS INSTRUCTION TIME.
M 0            + C29 SMALL GROUP INSTRUCTION TIME
  .#########   | C32 MOST ABLE GROUPED TOGETHER
               |
           .   | T24B [REV] STUDENT TIME IN CLASS WORK
    .#######   |
 -1        .   +
               |
S .##########  |
           #   |
               |
 -2  #######   +
               |
               |
Q       .#####
 -3            + C36 [REV] NO GROUPING
               |
     ########  |
               |
 -4    ######  +                                        <--- MORE COMMON
```

Figure 3

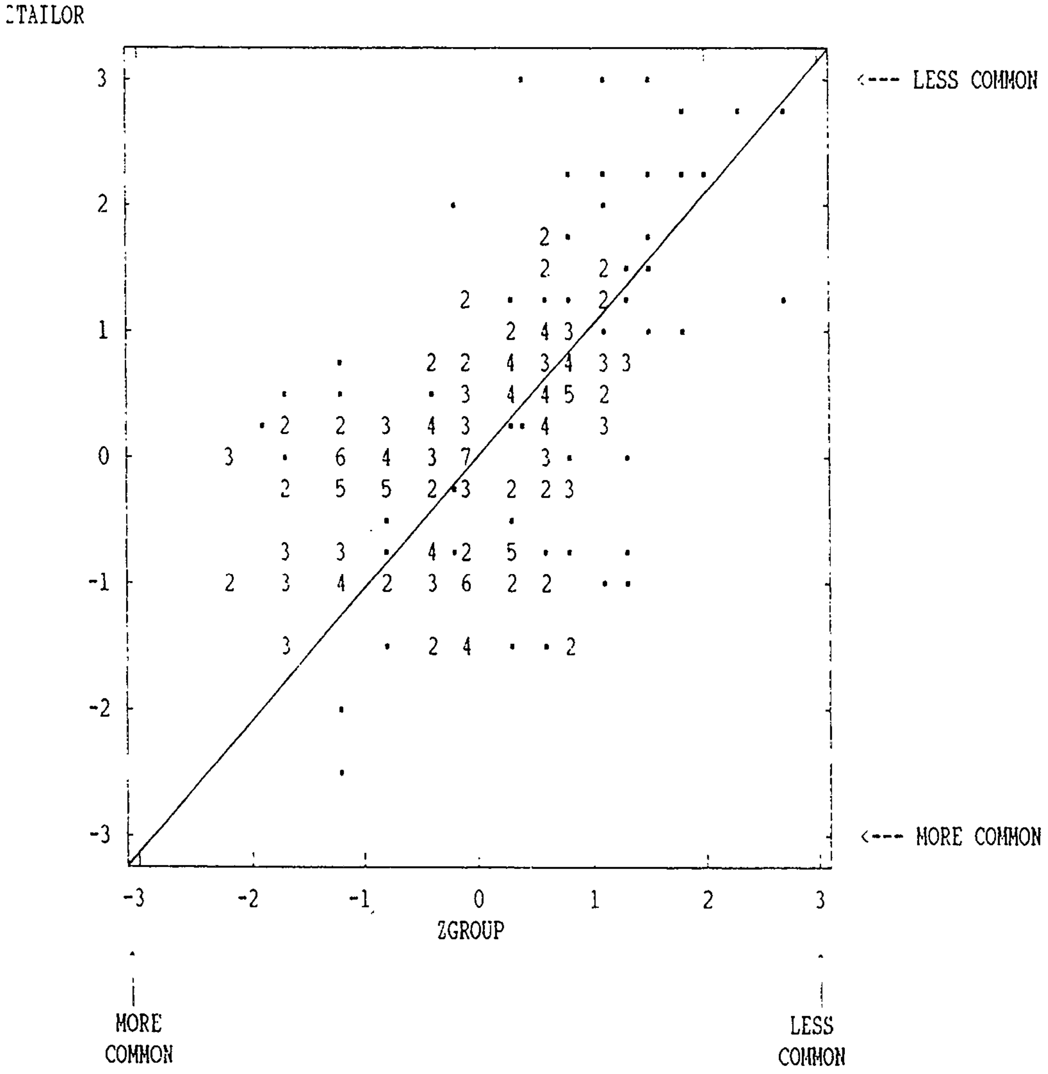
Person and Item Map for Tailoring Items

```
      PERSONS        MAP OF ITEMS

   4           # +                                                   <--- LESS COMMON
                 |
               # |
                 |
Q  3             +
                 |
              .# |
               . |
                 |
   2           # +
               . | C39 SOME ASSIGNED MORE EXERCISES
S             ## | C37B CONTENT VARIED
                 |
               # | C41 SOME ASSIGNED HARDER TOPICS
   1          .# + C37A PACING VARIED
              ## | C38A DUE DATES VARIED                C38B ASSIGNMENTS VARIED
              .# |
             .## | C40 SOME ASSIGNED HARDER EXERCISES
             ### |
M  0         .## + T26 [REV] FREQUENCY OF DIFFERENT ASSIGNMENTS
             .## |
           .#### |
               . |
        .####### |
  -1             +
      .######### | C62 IMP: SIMPLER TASKS FOR LESS ABLE
               # |
S       .####### |
                 | C97 IMP: ASSIGNMENTS TAILORED TO NEEDS
  -2           . + C91 IMP: VERY HARD TASKS FOR TRULY ABLE
        .####### |
               . |
                 |
       .######## | C67 IMP: HARDER TASKS FOR MORE ABLE
Q -3             +
                 |
                 |
           .#### |
                 |
  -4           # +                                                   <--- MORE COMMON
```

Figure 4

Grouping vs Tailoring Person Measure Plot



APPENDIX A

Time Spent in Grouping Arrangements Items

SIMS Classroom Process Questionnaire: Estimate the amount of class time in a typical week which is devoted to each of the following (in percentages):

C28. The whole class working together as a single group (e.g., whole class lecture or discussion).
C29. Small group instruction (or some combination of small groups and students working individually).
C30. All students working individually (with or without individual help from teacher or teacher aide).

The percentage of time spent in each arrangement was summed. In some cases the total of what teachers reported was substantially more or less than 100%. Therefore the percentage of time spent in each arrangement was compared to the total percentage of time reported for the three arrangements. Four levels (categories 0-3) of time spent in various grouping arrangements were created: None (0%), Minimal (1-33%), Moderate (34-66%), and Predominant (67-100%).

SIMS Teacher Questionnaire: T24. Now estimate the average time per student spent by the target class on each of the following: (estimate of number of minutes spent in each activity in a typical week)

T24A. Doing seatwork or blackboard work (students preparing individual written answers to assigned exercises or problems).
T24B. Listening as a whole class to you give lectures or explanations.
T24C. Working in small groups.

The amount of time spent in all three activities was summed. In some cases the total time teachers reported was substantially more or less than the total amount of math instruction or class time (from another item on the survey). Therefore the percentage of time spent in each activity was compared to the total amount of time reported for the three activities. Four levels (categories 0-3) of time spent in various activities were created: None (0%), Minimal (1-33%), Moderate (34-66%), and Predominant (67-100%).

Types of Grouping Arrangements Items

SIMS Classroom Process Questionnaire: Which of the following situations occur regularly in your small group instruction (Check as many as apply.)

C32. Most able students work separately while the rest of the class works as a single group.
C33. Least able students work separately while the rest of the class works as a single group.
C34. The class is split into 3 or more groups each at a different ability level.

Yes coded "1"
No coded "0"

Two additional items were created to provide data for two items not included on data tape.

C35. None of the above occurs regularly [interpreted as mixed ability grouping used].
C36. Question does not apply--no small group instruction.

Data for C35 was created by identifying teachers who indicated they spent some time in small group instruction but did not indicate any of the above grouping situations (coded "1") and the rest of the teachers were coded "0". Data for C36 was created by identifying teachers who indicated they spent no time in small group instruction (coded "1") and the rest of the teachers were coded "0".

Tailoring Belief Items

SIMS Classroom Process Questionnaire: Below you will find suggestions of what teachers might do to make their teaching more effective. Please rate each item as if you were selecting a shorter list of the more important items to emphasize with student teachers and others who are interested in effective teaching. Circle the appropriate number for each item as follows:

4 Among the highest in importance
3 Of major importance
2 Of some importance
1 Of little or no importance

C62. Give less able students assignments that are simple enough that they can progress without making mistakes.
C67. Assign problems which require the abler students to do more than follow examples that have already been demonstrated.
C73. Vary the difficulty of questions posed in classroom discussion.
C91. Give abler students assignments with some problems which are truly difficult for them to solve.
C97. Give assignments which are tailored to the particular instructional needs of individual students.

Tailoring Practice Items

SIMS Teacher Questionnaire: T26. How often are some students in the target class asked to do exercises or problem assignments which are different from those given other students in the class? (Check one)

3 Rarely or never (Scale reversed so that most frequent had highest value)
2 Occasionally
1 Frequently

SIMS Classroom Process Questionnaire: Which of the following statements best describes/is most characteristic of your class.

C37B. To the extent possible, I teach all students same content but let them proceed at their own pace.
C37C. To the extent possible, I vary the content across students or groups of students.
C38B. All students are assigned the same set of exercises but the date of completion varies from student to student.
C38C. Some students are assigned exercises that I would not expect other students in the class to do.

Yes coded "1"
No coded "0"

SIMS Classroom Process Questionnaire: To show how the exercises assigned some students differ from those assigned to other students, check those statements which are typical of your class.

C39. Some students are assigned more exercises than other students.
C40. Some students are assigned more difficult exercises than other students.
C41. Some students are assigned exercises on topics which other students aren't expected to cover this year.

Yes coded "1"
No coded "0"