Vol. 2 (1), pp. 3-21, Jan, 2015
http://www.ijate.net
e-ISSN: 2148-7456 © IJATE
INTERNATIONAL JOURNAL of ASSESSMENT TOOLS in EDUCATON

Quality Control in Survey Design: Evaluating a Survey of Educators' Attitudes Concerning Differentiated Compensation

Kelly D. BRADLEY[1,a], Michael PEABODY[b], Shannon O. SAMPSON[c]

[a]University of Kentucky, College of Education Department of Educational Policy Studies & Evaluation, 144A Taylor Education Building, Lexington, KY 40506-001, United States

[b]American Board of Family Medicine, United States

[c]Independent Consultant, United States

Abstract

This study utilized the Rasch model to assess the quality of a survey instrument designed to measure attitudes of administrators and teachers concerning a differentiated teacher compensation program piloted in Kentucky. Researchers addressing potentially contentious issues should ensure their methods stand up to rigorous criticism. The results indicate that the rating scale does not function as expected, with items being too easy to endorse. Future iterations of this survey should be revised prior to release. Recommendations for improvement are provided.

Keywords: Rasch measurement, Survey design, Rating Scale, Teacher Compensation

Article Info

Received
25 October 2015

Accepted:
11 December 2014

Keywords: Rasch measurement; Survey design; Rating Scale; Teacher Compensation

1. Introduction

Surveys are one of the most common examples of self-reported data collection and continue to be an ever-popular research methodology for graduate studies and published papers in education. However, the development of instruments to assess affective domain constructs has been a problematic area within the field of education. The quality of the instrument used in the measurement process must play a fundamental role in the analysis of the data collected from it. As Bond and Fox (2001) note,

> "Operationalizing and then measuring variables are two of the necessary first steps in the empirical research process. Statistical analysis, as a tool for investigating relations among the measures, then follows. Thus, interpretation of analyses can only be as good as the quality of the measures." (p.xvi)

It is therefore important to begin at the level of measurement and to identify weaknesses that may limit the reliability and validity of the measures made with a survey instrument. Only after the efficiency and effectiveness of the measurement instrument has been addressed should statistical analysis with the data take place.

[1]Corresponding Author Email: kdbrad2@uky.edu

Some psychometricians and behavioral statisticians treat survey data as if the mere assignment of numerical values to objects suffices as scientific measurement. Certain assumptions are put in place when researchers develop a group of items intended to assess a construct, administer the items to a sample of respondents, and sum the ratings. In such cases, researchers are assuming that (1) each item contributes equally to the measure of that construct, implying all items are of equal importance, (2) each item is measured on the same interval scale; and (3) respondents have appropriately interpreted the directions, all items are written clearly, and the items tap the same construct, creating a single dimension. In actuality, these assumptions are unstable and problematic in survey research methods (Hambleton, Swaminathan and Rogers, 1991; Becker, 2001; Bond, et al., 2001; Bradley and Sampson, 2005).

This study utilizes the Rasch model to assess the quality of the measurement instrument, here a selected response survey, and the structure of the rating scale. Findings will benefit the developers of the survey in ensuring the instrument is functioning as intended and will give insight into possible revisions for future rounds of data collection. In addition, a methodological framework for educational researchers developing survey instruments and analyzing rating scale data is offered.

2. Background

2.1 Rasch versus Classical Test Theory Approach

Researchers often utilize the classical test theory model in analyzing the rating scale data produced via the selected-response survey; however, classical test theory model, also referred to as the true score model, has deficiencies. The classical test theory approach requires complete records to make comparisons of items on the survey. Even if a dataset of complete responses is attained, the issues of sample-dependence between estimates of an item's difficulty to endorse and a respondent's willingness to endorse surface. Issues of sample-dependence are problematic as they make the estimates for item calibrations dependent on the severity of the respondents in the sample. Moreover, the estimates of item difficulty cannot be directly compared unless the estimates come from the same sample or assumptions are made about the comparability of the different samples. Another concern with the classical test theory approach is that only a single standard error of measurement is produced for the composite of the ratings or scores, making it inadequate and potentially misleading.

In this study, a methodological framework is presented to address the concerns related to the traditional classical test theory approach presented above. The quality of the instrument used to collect the data is first assessed; then, the collected data are analyzed and interpreted in a mathematically sound manner. Thus, the study presents a method for ensuring and/or establishing the quality of the survey instrument.

The Rasch model, introduced by Georg Rasch (1960), addresses many of the weaknesses of the classical test theory approach. The Rasch model allows for the connection of observations of respondents and items in a way that indicates the occurrence of a certain response as probability rather than certainty and maintains order such that the probability of providing a certain response defines an order of respondents and items. These circumstances create the probabilistic version of the scalogram (Guttman, 1944), which indicates a person endorsing a more extreme statement should also endorse all less extreme statements and an easy-to-endorse item is always expected to be rated higher by any respondent (Wright and Masters, 1982). Applying the Rasch model allows researchers to identify where possible misinterpretation occurs in the instrument and which items do not appear to measure the construct of interest. Information is also produced concerning the structure of the rating scale and the degree to which each item contributes to the construct. The model provides a mathematically sound alternative to traditional approaches of survey analysis.

In contrast to classical test theory, parameters in the Rasch model are neither sample nor test dependent, so missing data are not problematic (Rasch, 1960). As well, the Rasch approach produces standard error estimates for each discrete raw score, allowing for one reliability coefficient to be calculated for the instrument and another for the respondents. It is also possible to combine any person's estimated measure with any item's estimated measure to produce expected response value because respondents and items are measured on the same scale. As previously mentioned, applications of the Rasch model also provide estimates for persons and items that are freed from the sampling distribution of the sample employed, meaning there is no dependence on the particulars of the questionnaire or of the sample being measured (Wright and Masters, 1982). It is important that quantitative methodologist within the human sciences discontinue the practice of analyzing raw data or counts and instead analyze measures.

The analysis of measures is central to the Rasch model. This study applies Rasch techniques to a survey designed to measure the attitudes of teachers, mentor teacher-achievement coaches, superintendents, and principals toward differentiated compensation programs, defined as a range of incentives that are added on to present compensation. Such compensations include salary bonuses for teaching in critical shortage areas; financial support for seeking advanced degrees; or participation in voluntary career advancement opportunities. By attempting to measure the variable, it is understood that although respondents have many significant and distinct characteristics, only one characteristic can be meaningfully rated at a time.

2.2 Evaluating the Differentiated Teacher Compensation Program

The Kentucky Department of Education (KDE) provided support to ten school districts to plan and pilot a differentiated teacher compensation program to determine policy implications for (1) recruiting and retaining teachers in critical shortage areas; (2) reducing the number of emergency certified teachers; (3) providing incentives for teachers to serve in difficult assignments and hard-to-fill positions; (4) providing voluntary career advancement opportunities; and (5) rewarding teachers who increase their knowledge and skills. An essential component of the program is to ascertain the effectiveness of the program over a two-year period (2003-2005). KDE appointed the College of Education at the University of Kentucky (UK) to evaluate the Differentiated Compensation Pilot Site program[2]. A team of UK faculty constructed a selected response pencil-and-paper survey instrument as the essential measure, making it a key tool in the evaluation procedure. Items are outlined below.

Following the administration of the survey, a Rasch measurement approach was applied to the response data set in order to assess the stability and quality of the measures. Assumptions specific to the Rasch model guided the process of determining the quality of the set of measures related to the differentiated compensation survey. Specifically, the analysis responded to the questions of whether the survey measured a single variable and whether each person's response pattern indicated that they were responding in an acceptably predictable way given the expected hierarchy of responses (Wright and Masters, 1982). The results produce measures that adequately represent the intended construct with a replicable and meaningful set of measures. Findings have an immediate impact to the survey development team, while the methodology offers a framework for others, especially those in the social and behavioral sciences.

[2] The Principal Investigator in the study was Lars Björk, College of Education, University of Kentucky

3. Method

3.1. Participants

Surveys were distributed to school districts across Kentucky that had participated in the differentiated compensation pilot. Utilizing a census sample method this survey was administered to four groups: superintendents (n = 10), principals (n = 63), teachers (n = 438), and mentor teacher-achievement coaches (n = 60). The survey was not identical for each group due to the inherent differences of each group; however, each form of the survey was analogous. Because of the nature and distribution format of the surveys, the number of surveys distributed was approximately equivalent to the number of surveys returned. Thus, representativeness of the sample on the population is not a concern.

3.2 Apparatus

The selected response 34-item pencil-and-paper survey, along with the subsequent data collection tools, were developed and conducted by the group of researchers at the University of Kentucky. The items included topics ranging from the effects of differentiated compensation on factors such as standardized test scores, staff relations, and teacher morale and teacher recruitment; to teacher pride, identification and ownership in the school. Respondents were asked to rate their attitude toward differentiated compensation using a 4-point Likert-type scale labeled (4) “Agree”, (3) “Tend to Agree”, (2) “Tend to Disagree”, and (1) “Disagree”. One form of the survey was administered to the teachers and mentor teacher-achievement coaches and an analogous, but not identical, survey was administered to the principals and superintendents. Principals and superintendents were pooled to form an administrator subgroup and an assessment of the responses indicated the pooled administrator group (n = 73) was indeed homogeneous. A pooling of the two teacher subgroups was not performed because the mentor teacher-achievement coaches were determined to be positioned between teachers and administrators. Thus, Rasch analysis for this study was conducted for three groups: pooled administrators, teachers, and mentor teacher-achievement coaches.

3.3 Procedure

A one-parameter Item Response Theory model, commonly known as the Rasch model, was employed utilizing Winsteps Rasch Measurement Software (Linacre 2004). George Rasch (1960) described a dichotomous statistical measurement model to analyze test scores. The Rating Scale Model (Andrich, 1978; Wright and Masters, 1982), an extension of Rasch’s original dichotomous model designed to analyze ordered response categorical data, may be utilized to analyze Likert-type survey responses such as those in this study. The Rating Scale Model is shown here as:

$$\ln\left(\frac{Pnij}{Pni(j-1)}\right) = Bn - Di - Fj$$

Where $Pnij$ is the probability that person n encountering item i would be observed in category j; $Pni(j-1)$ is the probability that the observation would be in category $j-1$; Bn is the “ability” of person n; Di is the difficulty of item i; and Fj is the point where categories $j-1$ and j are equally probable relative to the measure of the item.

Winsteps utilizes the Andrich rating scale model with the Joint Maximum Likelihood Estimation method, also known as UCON, which does not assume a person distribution and is flexible with missing data (Wright and Masters, 1982). The Rasch model uses the sum of the item ratings simply as a starting point for estimating probabilities of those responding and because it is based upon the ability to endorse a set of items and the difficulty of a set of items, it is assumed item difficulty is the main characteristic influencing responses. Here, two facets are involved, the instrument's items and the respondents. From a Rasch perspective, a respondent's willingness to endorse interacts with an item's difficulty to assign a certain score to produce an observed outcome. In general, people are more likely to endorse easy-to-endorse items than those that are difficult to endorse, and people with higher willingness-to-endorse scores are more agreeable than those with low scores.

Rasch analysis reports person willingness-to-endorse and item difficulty-to-endorse estimates along a logit (log odds unit) scale. A logit scale is, "a unit interval scale in which the unit intervals between the locations on the person-item map have a consistent value or meaning" (Bond and Fox, 2001). Rasch measurement uses a logarithmic transformation of the item and person data to convert ordinal-level responses into interval-level data.

Rasch measurement establishes the model a priori and produces fit statistics to examine how well the data fit the model, as opposed to modeling the data. A review of an instrument begins with an examination of fit statistics for the survey items and respondents (Table 1). This study utilized ZSTD scores as the primary fit statistic. ZSTD scores are mean-square fit statistics standardized to approximate a theoretical mean 0 and standard deviation 1. INFIT ZSTD scores are sensitive to irregular inlying patterns and OUTFIT ZSTD scores are sensitive to unexpected rare extremes (Linacre, 2002b). Survey items and respondents that did not adequately fit the model requirements were identified using the ZSTD scores, with a cutoff set at 2. While there is not a specific rule defining the cutoff, the commonly accepted interpretation is that INFIT and OUTFIT values greater than +2 or less than –2 indicate less compatibility with the model than expected (Bond and Fox, 2001). The fit statistics are used to mark items for further scrutiny with the recommendation that certain items be rewritten or excluded for future analyses instead of blindly interpreting the total raw score for all persons on all items as the total construct measure.

Table 1. Summary Statistics

	Person			Item		
	Mean	S.D.	Reliability	Mean	S.D.	Reliability
Administrators			0.87			0.95
Measure	1.74	0.95		0	1.07	
Infit ZSTD	0	1.6		-0.1	1.8	
Outfit ZSTD	-0.1	1.6		-0.1	1.8	
Teacher-Mentor Coaches			0.83			0.96
Measure	1.4	0.71		0	1.34	
Infit ZSTD	-0.1	1.4		0	1.5	
Outfit ZSTD	0	1.6		0.1	1.6	
Teachers			0.84			0.99
Measure	1.24	0.75		0	1.01	
Infit ZSTD	-0.1	1.7		0	3.9	
Outfit ZSTD	0	1.6		0.1	4.3	

Further diagnostic statistics were reviewed as a way to understand, assess and suggest improvements to the measurement, or data collection, system. The output includes (1) item polarity; (2) empirical item-category measures; (3) category function; and (4) a graph of probability modes. Item polarity is a point-measure correlation whereby properly functioning items should exhibit a positive correlation coefficient. Empirical item-category measures provide a visual display of whether category values are properly ordered and increase in a 1-2-3-4 fashion. Category function measures allow the researcher to determine the level to which categories advance and the graph of probability modes is a visual representation of the category function measures. At some point along the continuum each category should be the most probable as shown by a distinct peak on the graph. Linacre (2002a) suggested that categories should advance by at least 1 logit but not more than 5 logits.

A person-item maps set respondents and survey items together on the same scale so instrument developers can clearly identify which items are more difficult to endorse and which respondents are more agreeable. Gaps between items suggest that items could be added to better span measure all points along the construct. Since person-item maps show both respondents and items on the same scale it is possible to more accurately target items to the overall agreeability of the sample. For example, if the map revealed the mass of the survey items above the mass of the respondents, it would indicate the questions were overall difficult to endorse for the corresponding sample. In this example, the instrument developer may choose to add new items that are easier to endorse and would be positioned lower on the item hierarchy. Matching the items to the respondents allows for more accurate measurement along the scale.

The initial Rasch analyses were run using the data as they were coded upon collection. A review of the item polarity identified certain items having negative point-measure correlations across most of the groups. The recoded items were:

- Item 2: “Differentiated compensation would not enhance the positive relationship among teachers and administrators”;
- Item 3: “A differentiated compensation program will have a negative impact on the morale of teachers in the system”;
- Item 7: “Linking teacher salary to student achievement on standardized tests has no place in education”;
- Item 8: “Teachers receiving differentiated compensation will be less cooperative with their peers”; and
- Item 12: “Relations between administrative and instructional staff will be negatively affected if a differentiated compensation program is adopted”.

Since the wording of these items encouraged a reverse use of the scale by respondents, these items were reverse-coded in the Rasch analyses to compensate for the negative wording and improve the diagnostics. Item 17 (“Non-certified teachers should pay all the costs of becoming certified”), was negatively correlated for some groups and had high outfit ZSTD scores prior to recoding, so it was recoded as well. Still, it could be argued that this item is better left with the original coding. Item 33 (“There is too much peer pressure here to do a good job”), was not reverse coded but could be viewed as a negative statement by some respondents. Thus, items 17 and 33 are two items that should be reviewed by the survey construction team. The item polarity point-measure correlation coefficients for each item following recoding are found in Table 2.

Table 2. Item Polarity Point Measure Correlation

Entry Number	Administrators	Mentor-Coach	Teachers	Entry Number	Administrators	Mentor-Coach	Teachers
1	0.6	0.54	0.62	18	0.29	0.54	0.38
2	0.37	0.56	0.61	19	0	-0.03	0.15
3	0.52	0.69	0.62	20	0.73	0.66	0.71
4	0.62	0.52	0.61	21	0.66	0.54	0.62
5	0.65	0.77	0.71	22	0.67	0.57	0.6
6	0.71	0.59	0.67	23	0.67	0.58	0.63
7	0.37	0.16	0.3	24	0.73	0.43	0.53
8	0.65	0.28	0.45	25	0.57	0.4	0.35
9	0.42	0.14	0.28	26	0.65	0.29	0.5
10	0.57	0.62	0.62	27	0.38	0.23	0.43
11	0.65	0.47	0.6	28	0.49	0.23	0.38
12	0.64	0.61	0.56	29	0.47	0.24	0.32
13	0.42	0.17	0.28	30	0.53	0.17	0.38
14	0.52	0.37	0.35	31	0.42	0.14	0.35
15	0.64	0.69	0.67	32	0.32	0.16	0.29
16	0	0	0.07	33	0.22	0.02	-0.16
17	0.2	0.44	0.21	34	0.19	0.21	0.33

4. Results

4.1. Administrators

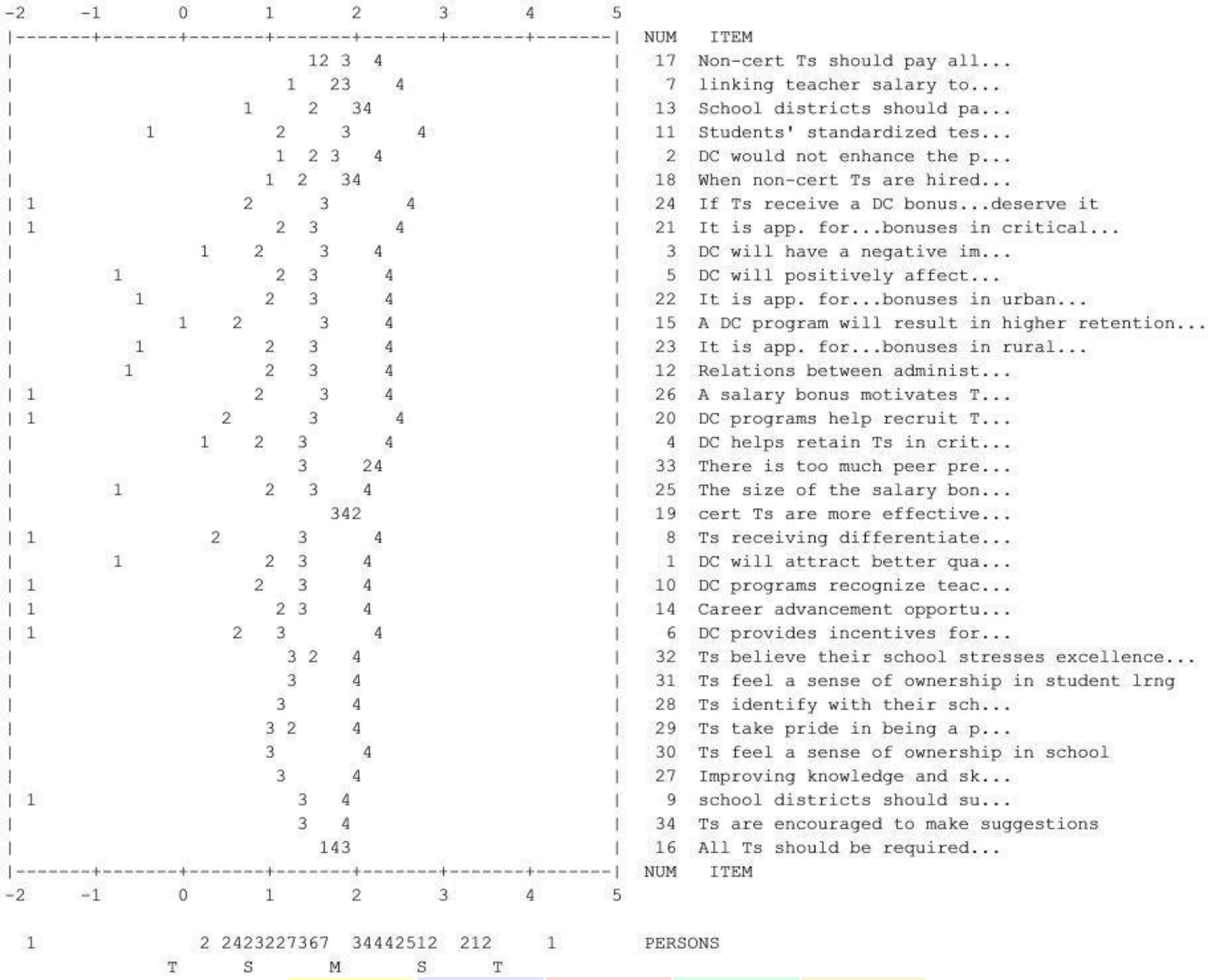
The survey data collected from the pooled group of administrators indicates a stable measurement instrument. The summary statistics (Table 1) reveal an acceptable person reliability of .87 and item reliability of .95. The Rasch reliability is a correlation coefficient, the ratio of true measure variance to observed measure variance, similar to a KR-20 or Cronbach Alpha (Fisher,1992; Wright, 1996). Thus, with these reliability estimates, it is reasonable to assume that another group of respondents with a similar average agreeability measure would produce similar results. For the administrators group, all items have positive correlations as displayed in the Item Polarity output (Table 2); however, the empirical item-category measures output (Figure 1) reveals certain items' category values do not function with category values corresponding to "more" of the next variable. Items whose values do not increase as expected include items 16, 19, 29, 32, and 33.

As displayed in Table 3, average measures for the rating scale categories do advance, and no category is especially noisy. Still, categories 3 and 4 on the scale comprise 87% of the total responses, illustrating the 4-point scale is functioning as a nearly dichotomous scale. The category probabilities graph (Figure 2) reveals that category 2 is almost flat, when it is expected that each category should have a distinct peak. This finding reveals that the "Tend to Disagree" choice does not aid in defining a distinct point on the variable. Finally, the person-item map for the pooled administrators (Figure 3) reveals the mean measure for the survey items is well below the mean measure for the respondents, indicating the items are generally easy to endorse for the administrators, which are agreeable to the constructed items.

Table 3. Item-Category Measures Summary

Category Label	Administrators	Mentor Teachers	Teachers
1	-2.54	-2.36	-2.2
2	-0.96	-0.78	-0.7
3	0.7	0.65	0.6
4	2.95	2.57	2.37

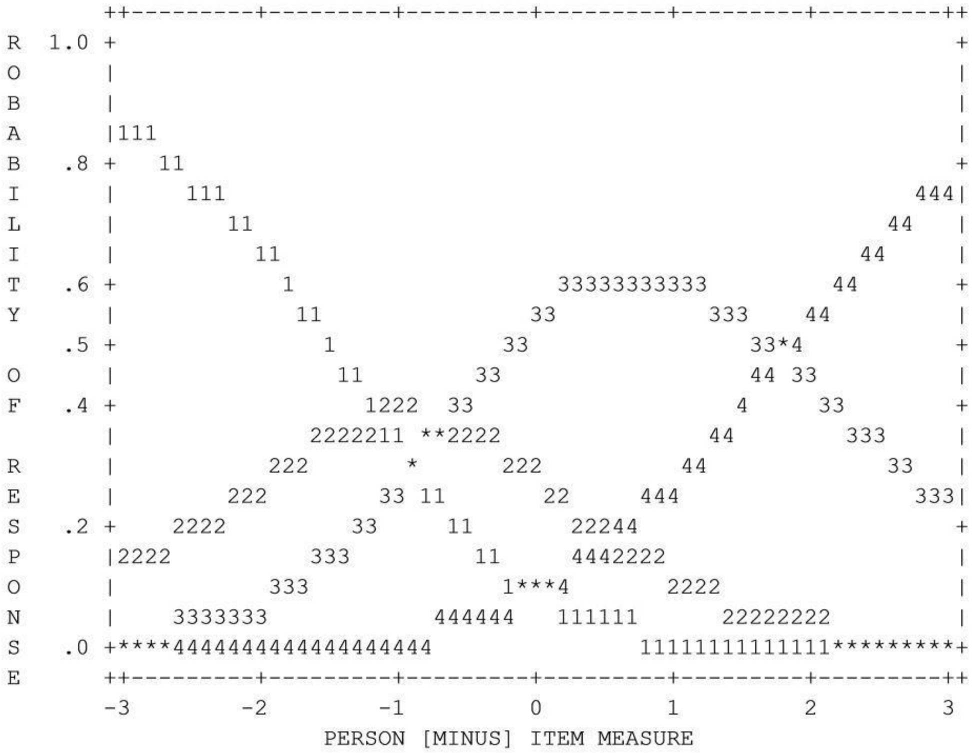
For the group of pooled administrators, no category has extremely large infit or outfit ZSTD scores (Table 4). The largest value reported is for item 17 (“Non-certified teachers should pay all the costs of becoming certified”), which has an outfit ZSTD of 4.0. This is followed by item 19 (“Certified teachers are more effective than non-certified teachers”), with an outfit ZSTD of 3.0. Five items have a negative outfit ZSTD greater than the traditional cutoff of –2, but none is less than –3.5. Negative outfit ZSTD suggests less variation than the probabilistic model would predict. It is not surprising their responses might be predictable given this is a group of administrators within the same state, sharing common issues.



```
-2     -1       0       1       2       3       4       5
|-------+-------+-------+-------+-------+-------+-------| NUM   ITEM
|                           12 3  4                     |  17  Non-cert Ts should pay all...
|                         1   23    4                   |   7  linking teacher salary to...
|                     1     2   34                      |  13  School districts should pa...
|            1           2     3      4                 |  11  Students' standardized tes...
|                        1  2 3   4                     |   2  DC would not enhance the p...
|                       1  2   34                       |  18  When non-cert Ts are hired...
| 1                   2      3       4                  |  24  If Ts receive a DC bonus...deserve it
| 1                      2  3       4                   |  21  It is app. for...bonuses in critical...
|                 1    2     3    4                     |   3  DC will have a negative im...
|         1              2  3      4                    |   5  DC will positively affect...
|           1           2   3      4                    |  22  It is app. for...bonuses in urban...
|               1    2       3     4                    |  15  A DC program will result in higher retention...
|           1           2   3      4                    |  23  It is app. for...bonuses in rural...
|          1            2   3      4                    |  12  Relations between administ...
| 1                    2     3     4                    |  26  A salary bonus motivates T...
| 1                 2       3       4                   |  20  DC programs help recruit T...
|                 1    2   3       4                    |   4  DC helps retain Ts in crit...
|                          3     24                     |  33  There is too much peer pre...
|         1             2   3    4                      |  25  The size of the salary bon...
|                             342                       |  19  cert Ts are more effective...
| 1                2       3      4                     |   8  Ts receiving differentiate...
|         1             2  3     4                      |   1  DC will attract better qua...
| 1                    2   3     4                      |  10  DC programs recognize teac...
| 1                      2 3     4                      |  14  Career advancement opportu...
| 1                  2   3        4                     |   6  DC provides incentives for...
|                         3 2   4                       |  32  Ts believe their school stresses excellence...
|                         3     4                       |  31  Ts feel a sense of ownership in student lrng
|                        3      4                       |  28  Ts identify with their sch...
|                       3 2     4                       |  29  Ts take pride in being a p...
|                       3        4                      |  30  Ts feel a sense of ownership in school
|                        3      4                       |  27  Improving knowledge and sk...
| 1                        3   4                        |   9  school districts should su...
|                          3   4                        |  34  Ts are encouraged to make suggestions
|                            143                        |  16  All Ts should be required...
|-------+-------+-------+-------+-------+-------+-------| NUM   ITEM
-2     -1       0       1       2       3       4       5

  1               2 2423227367  34442512  212     1       PERSONS
               T      S       M       S      T
```

Figure 1. Item-Category Measures for Administrators

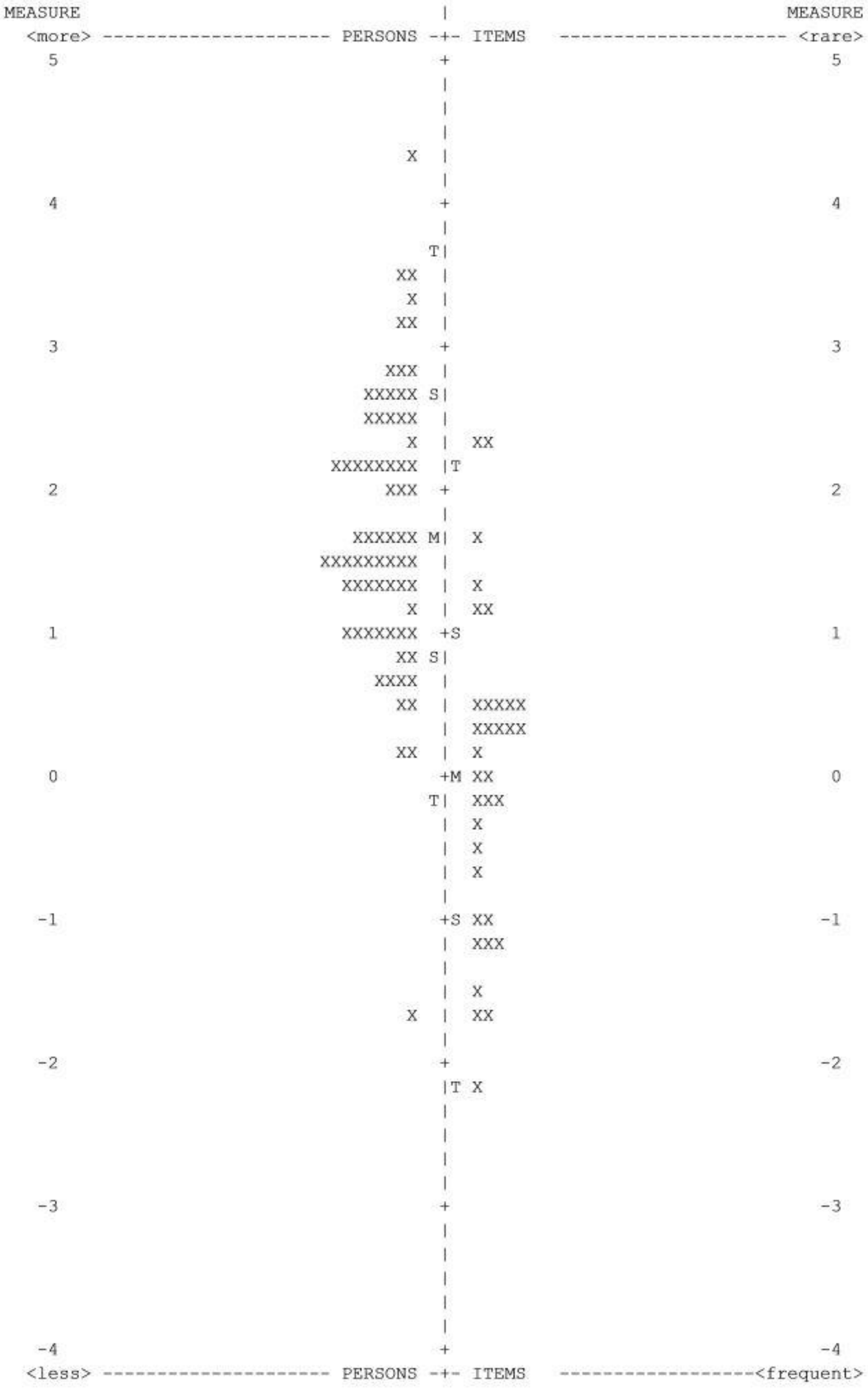


```
       ++---------+---------+---------+---------+---------+---------++
R  1.0 +                                                             +
O      |                                                             |
B      |                                                             |
A      |111                                                          |
B   .8 +   11                                                        +
I      |     111                                                  444|
L      |        11                                              44   |
I      |          11                                          44     |
T   .6 +            1                   33333333333         44       +
Y      |             11               33           333    44         |
    .5 +               1            33                33*4           +
O      |                11        33                  44 33          |
F   .4 +                  1222  33                   4     33        +
       |              2222211 **2222               44        333     |
R      |           222       *      222          44             33   |
E      |        222        33 11       22     444                 333|
S   .2 +    2222         33     11       22244                       +
P      |2222          333         11     4442222                     |
O      |           333              1***4       2222                 |
N      |    3333333            444444   111111      22222222         |
S   .0 +****4444444444444444444               11111111111111*********+
E      ++---------+---------+---------+---------+---------+---------++
       -3        -2        -1         0         1         2         3
                       PERSON [MINUS] ITEM MEASURE
```

Figure 2. Category Probability Curves for Administrators

Table 4. Items by misfit for administrators

Entry Number	Infit ZSTD	Outfit ZSTD
17	4.1	4
19	2.1	3
2	2.6	2.8
7	2.6	2.6
16	2.9	2.5
18	2.4	2.5
8	-2	-2.1
6	-2.1	-2.5
24	-3	-2.9
20	-2.9	-3.1
11	-3.8	-3.5



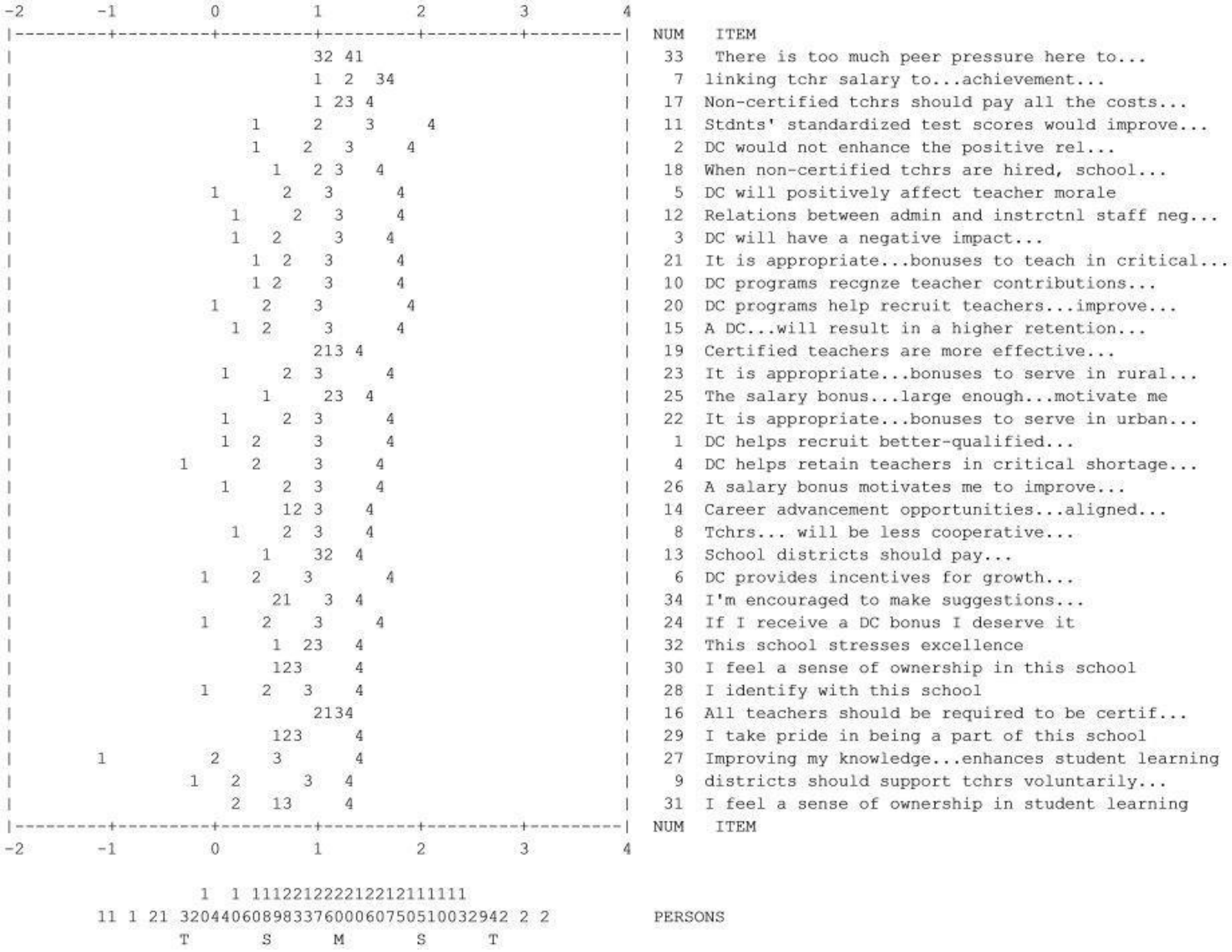
```
MEASURE                                 |                              MEASURE
  <more> --------------------- PERSONS -+- ITEMS  --------------------- <rare>
    5                                   +                                   5
                                        |
                                        |
                                        |
                                     X  |
                                        |
    4                                   +                                   4
                                        |
                                       T|
                                    XX  |
                                     X  |
                                    XX  |
    3                                   +                                   3
                                   XXX  |
                                 XXXXX S|
                                 XXXXX  |
                                     X  |  XX
                              XXXXXXXX  |T
    2                              XXX  +                                   2
                                        |
                                XXXXXX M|  X
                             XXXXXXXXX  |
                               XXXXXXX  |  X
                                     X  |  XX
    1                          XXXXXXX  +S                                  1
                                    XX S|
                                  XXXX  |
                                    XX  |  XXXXX
                                        |  XXXXX
                                    XX  |  X
    0                                   +M XX                               0
                                       T|  XXX
                                        |  X
                                        |  X
                                        |  X
                                        |
   -1                                   +S XX                              -1
                                        |  XXX
                                        |
                                        |  X
                                     X  |  XX
                                        |
   -2                                   +                                  -2
                                        |T X
                                        |
                                        |
                                        |
                                        |
   -3                                   +                                  -3
                                        |
                                        |
                                        |
                                        |
                                        |
   -4                                   +                                  -4
  <less> --------------------- PERSONS -+- ITEMS  ------------------<frequent>
```

Figure 3. Item-Person Map for Administrators

4.2. Teachers

The Rasch analysis for the teacher survey reveals a less effective measurement system or survey instrument. Summary statistics presented in Table 1 indicate teachers are generally less agreeable on the items than the administrators. The mean teacher measure is 1.24 logits with a standard deviation of .75, while the mean administrator measure is 1.74 logits with a standard deviation of .95. Person reliability is .84 and item reliability is high at .99, so the instrument is functioning acceptably overall, although there are concerns to be addressed.

The item polarity table (Table 2) indicates that item 33 stands apart as the only item with a negative point-measure correlation. Item 33 also appears on the empirical item-category measures chart (Figure 4) as having misordered category values of 2-3-4-1 rather than 1-2-3-4 as expected. Other items with misordered category values are 16, 19, and 31.



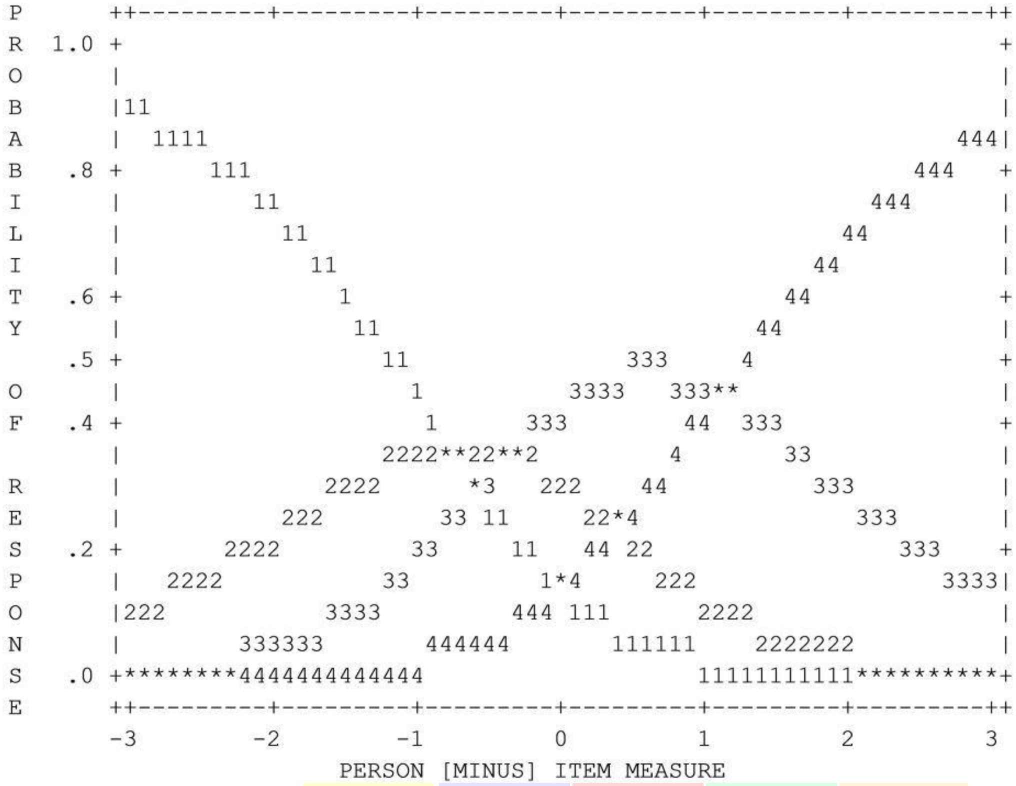
```
-2       -1         0         1         2         3         4
|---------+---------+---------+---------+---------+---------|  NUM   ITEM
|                             32 41                         |   33   There is too much peer pressure here to...
|                             1  2  34                      |    7  linking tchr salary to...achievement...
|                             1 23 4                        |   17  Non-certified tchrs should pay all the costs...
|                       1     2    3     4                  |   11  Stdnts' standardized test scores would improve...
|                       1    2   3     4                    |    2  DC would not enhance the positive rel...
|                         1   2 3   4                       |   18  When non-certified tchrs are hired, school...
|                   1      2   3      4                     |    5  DC will positively affect teacher morale
|                     1     2   3     4                     |   12  Relations between admin and instrctnl staff neg...
|                     1   2     3    4                      |    3  DC will have a negative impact...
|                       1  2   3      4                     |   21  It is appropriate...bonuses to teach in critical...
|                       1 2    3      4                     |   10  DC programs recgnze teacher contributions...
|                   1    2    3        4                    |   20  DC programs help recruit teachers...improve...
|                     1  2     3      4                     |   15  A DC...will result in a higher retention...
|                             213 4                         |   19  Certified teachers are more effective...
|                    1     2  3      4                      |   23  It is appropriate...bonuses to serve in rural...
|                        1     23  4                        |   25  The salary bonus...large enough...motivate me
|                    1     2  3      4                      |   22  It is appropriate...bonuses to serve in urban...
|                    1  2     3      4                      |    1  DC helps recruit better-qualified...
|                1      2     3     4                       |    4  DC helps retain teachers in critical shortage...
|                    1     2  3     4                       |   26  A salary bonus motivates me to improve...
|                          12 3    4                        |   14  Career advancement opportunities...aligned...
|                     1    2  3    4                        |    8  Tchrs... will be less cooperative...
|                        1    32  4                         |   13  School districts should pay...
|                  1    2    3       4                      |    6  DC provides incentives for growth...
|                         21   3  4                         |   34  I'm encouraged to make suggestions...
|                  1     2    3     4                       |   24  If I receive a DC bonus I deserve it
|                         1  23   4                         |   32  This school stresses excellence
|                         123     4                         |   30  I feel a sense of ownership in this school
|                  1     2   3    4                         |   28  I identify with this school
|                             2134                          |   16  All teachers should be required to be certif...
|                         123     4                         |   29  I take pride in being a part of this school
|        1          2     3       4                         |   27  Improving my knowledge...enhances student learning
|                 1   2      3   4                          |    9  districts should support tchrs voluntarily...
|                     2   13     4                          |   31  I feel a sense of ownership in student learning
|---------+---------+---------+---------+---------+---------|  NUM   ITEM
-2       -1         0         1         2         3         4

                   1  1 111221222212212111111
         11 1 21 32044060898337600060750510032942 2 2          PERSONS
                 T       S      M       S      T
```

Figure 4. Item-Category Measures for Teachers

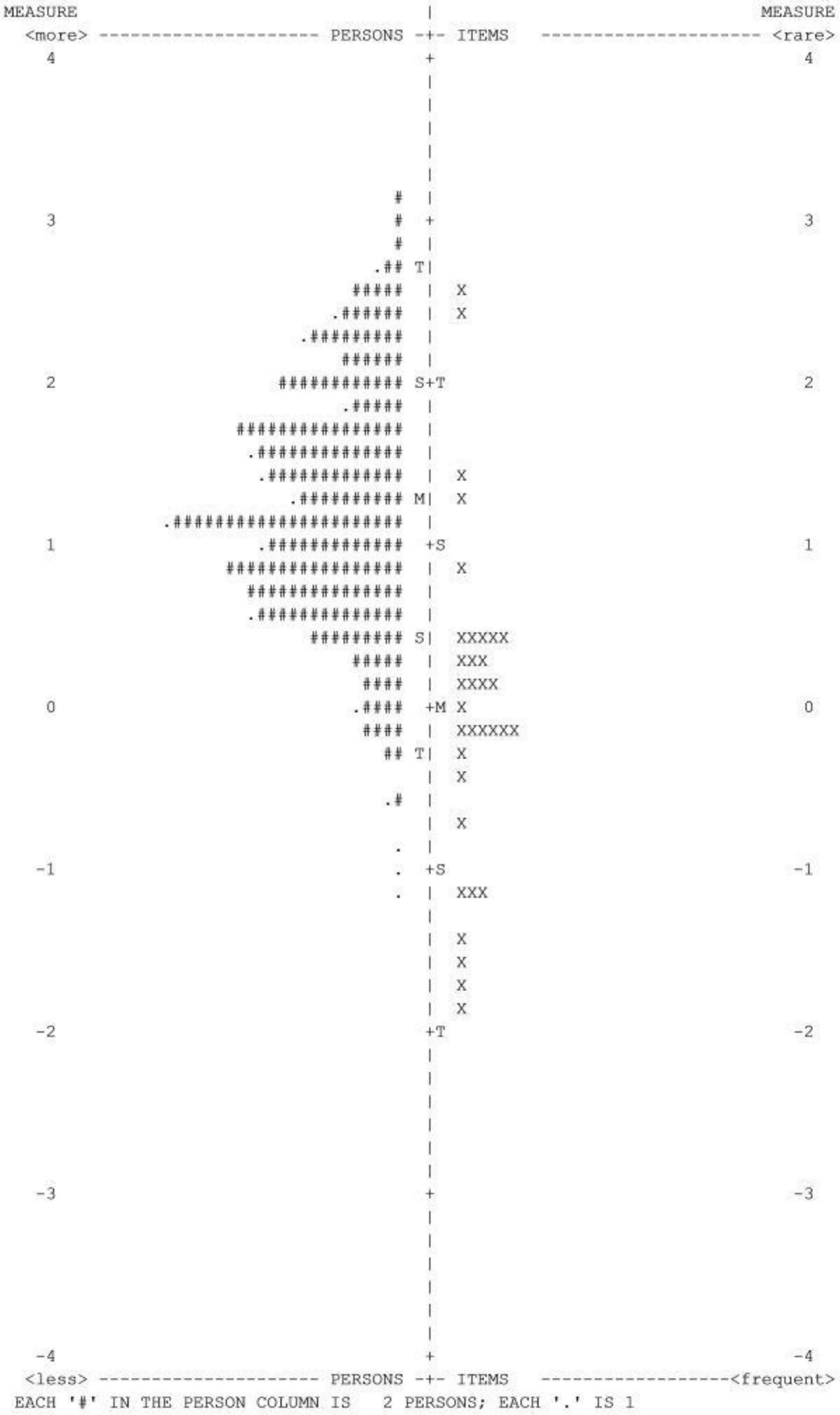
Overall, the category average measures advance, as displayed in Table 3, although respondents select 3 and 4 most often. Here again, category 2 is relatively flat, as shown in Figure 5. It is possible to collapse categories 1 and 2; this would result in probability curves that show each category represents a distinct portion of the variable and would likely improve the rating scale diagnostics. However, unless the survey is reconstructed to demonstrate such collapsing, it is not recommended, as interpretation of results might be altered, causing a misinterpretation of the original responses. As with the administrators, the person-item map for teachers (Figure 6) reveals the mean measure for the survey items is well below the mean measure for the respondents, indicating the items are generally easy to endorse for this group, which is agreeable to the constructed items.

There are multiple misfitting items for the teachers. Eleven items have an outfit ZSTD greater than 2 and as high as 9.9, and 14 items have outfit ZSTD less than –2. Items with high outfit ZSTD scores are highlighted in Table 5. Outfit ZSTD scores greater than 2 indicate unexpected and unrelated irregularities in the responses. There are many plausible causes for this, including ambiguous wording, negative wording, and debatable or misleading options. Twelve items have outfit ZSTD less than –2. Items with outfit ZSTD less than –2 include item 20 and item 11. As previously noted, large negative fit statistics indicate there is less variability than the model would predict. This can be caused by redundancy among items or could be a reflection of high agreement across respondents.



```
P      ++---------+---------+---------+---------+---------+---------++
R  1.0 +                                                             +
O      |                                                             |
B      |11                                                           |
A      |  1111                                                    444|
B   .8 +      111                                             444   +
I      |         11                                         444      |
L      |           11                                     44         |
I      |             11                                 44           |
T   .6 +               1                              44             +
Y      |                11                          44               |
    .5 +                  11               333     4                 +
O      |                    1          3333   333**                  |
F   .4 +                     1      333        44  333               +
       |                  2222**22**2         4       33             |
R      |              2222      *3   222    44          333          |
E      |           222        33 11     22*4               333       |
S   .2 +       2222         33     11   44 22                 333    +
P      |   2222           33         1*4     222                 3333|
O      |222           3333         444 111      2222                 |
N      |        333333       444444       111111    2222222          |
S   .0 +********444444444444                    1111111111***********+
E      ++---------+---------+---------+---------+---------+---------++
       -3        -2        -1         0         1         2         3
                       PERSON [MINUS] ITEM MEASURE
```

Figure 5. Category Probability Curves for Teachers



```
MEASURE                                 |                                MEASURE
  <more> --------------------- PERSONS -+- ITEMS    --------------------- <rare>
    4                                   +                                    4
                                        |
                                        |
                                        |
                                        |
                                        |
                                     #  |
    3                                #  +                                    3
                                     #  |
                                   .## T|
                                 #####  |  X
                               .######  |  X
                            .#########  |
                                ######  |
    2                     ############ S+T                                   2
                                .#####  |
                      ################  |
                       .##############  |
                        .#############  |  X
                           .########## M|  X
               .######################  |
    1                   .#############  +S                                   1
                     #################  |  X
                       ###############  |
                       .##############  |
                             ######### S|  XXXXX
                                 #####  |  XXX
                                  ####  |  XXXX
    0                            .####  +M X                                 0
                                  ####  |  XXXXXX
                                    ## T|  X
                                        |  X
                                    .#  |
                                        |  X
                                     .  |
   -1                                .  +S                                  -1
                                     .  |  XXX
                                        |
                                        |  X
                                        |  X
                                        |  X
                                        |  X
   -2                                   +T                                  -2
                                        |
                                        |
                                        |
                                        |
                                        |
                                        |
   -3                                   +                                   -3
                                        |
                                        |
                                        |
                                        |
                                        |
                                        |
   -4                                   +                                   -4
  <less> --------------------- PERSONS -+- ITEMS    ------------------<frequent>
 EACH '#' IN THE PERSON COLUMN IS   2 PERSONS; EACH '.' IS 1
```

Figure 6. Item-Person Map for Teachers

Table 5. Items by misfit for teachers

Entry Number	Infit ZSTD	Outfit ZSTD
33	7.2	9.6
17	7.3	8.6
19	5.9	7.3
16	4.9	5.8
13	5.3	5.6
7	4.8	5.4
18	2.9	4.2
32	3.1	3.9
34	3.7	3.4
25	3	3.4
22	-1.5	-2.3
10	-2.2	-2.3
23	-1.6	-2.4
27	-1.2	-2.6
2	-3.6	-2.7
3	-2.7	-3
1	-2.9	-3.5
26	-2.1	-3.6
15	-4.8	-3.6
4	-4.5	-4.6
5	-5.2	-5.1
6	-4.7	-5.3
20	-6.6	-6.4
11	-7.6	-7.3

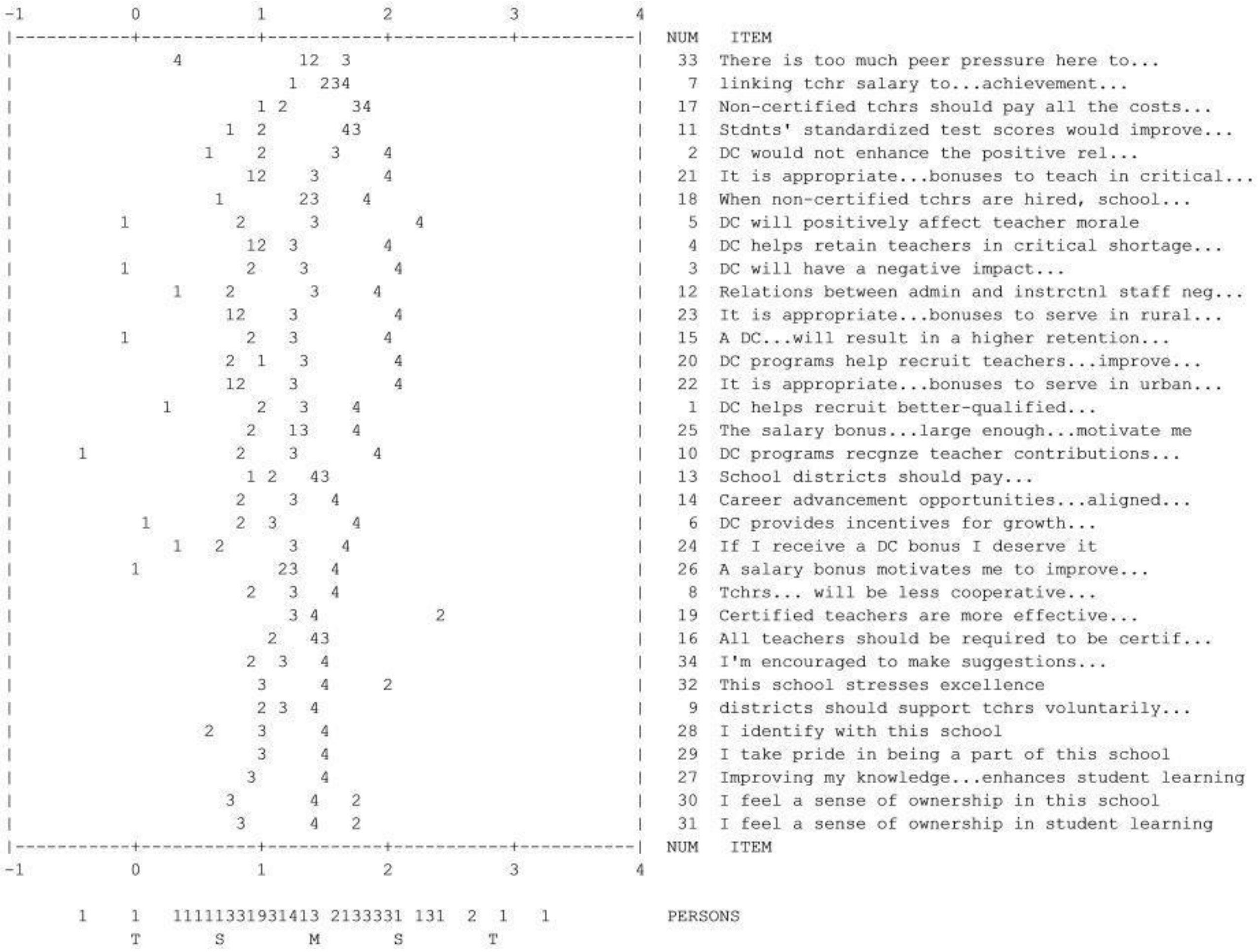
4.3. Mentor Teacher-Achievement Coaches

The instrument, again, functions well for the mentor teacher-achievement coach respondents, but discrepancies exist. Summary statistics displayed in Table 9 indicate the mentor teachers have a mean measure of 1.40 with a standard deviation of .71, placing between the teachers and the administrators, as previously hypothesized. Person reliability for this group is .83 and item reliability is .96; again reasonable measures.

The item-polarity table (Table 2) indicates that item 19 was the only item with a negative point-measure correlation. As illustrated by Figure 7, items with categories out of the expected 1-2-3-4 order are items 11, 13, 19, 20, 25, 30, 31, 32, and 33. As displayed in Table 3, average measures for the rating scale categories do advance and no category is especially noisy. Still, here again categories 3 and 4 on the scale dominate the responses, comprising 80% of the total responses and again illustrating that the 4-point scale is again functioning as a nearly dichotomous scale. Also, the category probabilities graph (Figure 8) reveals that category 2 is almost flat, when it is expected that each category should have a distinct peak.

Similar to both the administrators and teachers, the person-item map for mentor teacher-achievement coaches (Figure 9) reveals the mean measure for the survey items is well below the mean measure for the respondents. Here again, the items are generally easy to endorse for this group, which is agreeable to the constructed items.

Fit statistics for the mentor teacher-achievement coach group are much better than for the teachers or administrators (Table 6). Item 33 is again flagged as misfitting with an outfit ZSTD of 3.1. The other two items with an outfit ZSTD of greater than 2 for this group are items 13 and 19. There are fewer items with high negative outfit ZSTD scores as well with the largest being –3.6 for item 5. Other items with outfit ZSTD of less than –2 are items 3, 10, 12, and 20.



```
-1          0           1           2           3           4
|-----------+-----------+-----------+-----------+-----------|  NUM   ITEM
|               4           12  3                           |   33  There is too much peer pressure here to...
|                          1  234                           |    7  linking tchr salary to...achievement...
|                       1 2      34                         |   17  Non-certified tchrs should pay all the costs...
|                    1  2       43                          |   11  Stdnts' standardized test scores would improve...
|                  1    2      3    4                       |    2  DC would not enhance the positive rel...
|                      12    3      4                       |   21  It is appropriate...bonuses to teach in critical...
|                   1       23    4                         |   18  When non-certified tchrs are hired, school...
|          1          2      3         4                    |    5  DC will positively affect teacher morale
|                      12  3        4                       |    4  DC helps retain teachers in critical shortage...
|          1           2    3        4                      |    3  DC will have a negative impact...
|               1    2       3     4                        |   12  Relations between admin and instrctnl staff neg...
|                    12    3         4                      |   23  It is appropriate...bonuses to serve in rural...
|          1           2   3        4                       |   15  A DC...will result in a higher retention...
|                    2  1   3        4                      |   20  DC programs help recruit teachers...improve...
|                    12    3         4                      |   22  It is appropriate...bonuses to serve in urban...
|              1        2   3    4                          |    1  DC helps recruit better-qualified...
|                      2   13    4                          |   25  The salary bonus...large enough...motivate me
|      1              2    3       4                        |   10  DC programs recgnze teacher contributions...
|                      1 2   43                             |   13  School districts should pay...
|                     2    3   4                            |   14  Career advancement opportunities...aligned...
|            1        2  3       4                          |    6  DC provides incentives for growth...
|               1   2      3    4                           |   24  If I receive a DC bonus I deserve it
|           1             23   4                            |   26  A salary bonus motivates me to improve...
|                      2   3   4                            |    8  Tchrs... will be less cooperative...
|                          3 4           2                  |   19  Certified teachers are more effective...
|                        2   43                             |   16  All teachers should be required to be certif...
|                      2  3   4                             |   34  I'm encouraged to make suggestions...
|                       3     4     2                       |   32  This school stresses excellence
|                       2 3  4                              |    9  districts should support tchrs voluntarily...
|                  2    3     4                             |   28  I identify with this school
|                       3     4                             |   29  I take pride in being a part of this school
|                      3      4                             |   27  Improving my knowledge...enhances student learning
|                    3       4   2                          |   30  I feel a sense of ownership in this school
|                     3      4   2                          |   31  I feel a sense of ownership in student learning
|-----------+-----------+-----------+-----------+-----------|  NUM   ITEM
-1          0           1           2           3           4

       1    1   11111331931413 2133331 131  2  1   1           PERSONS
            T       S        M       S        T
```

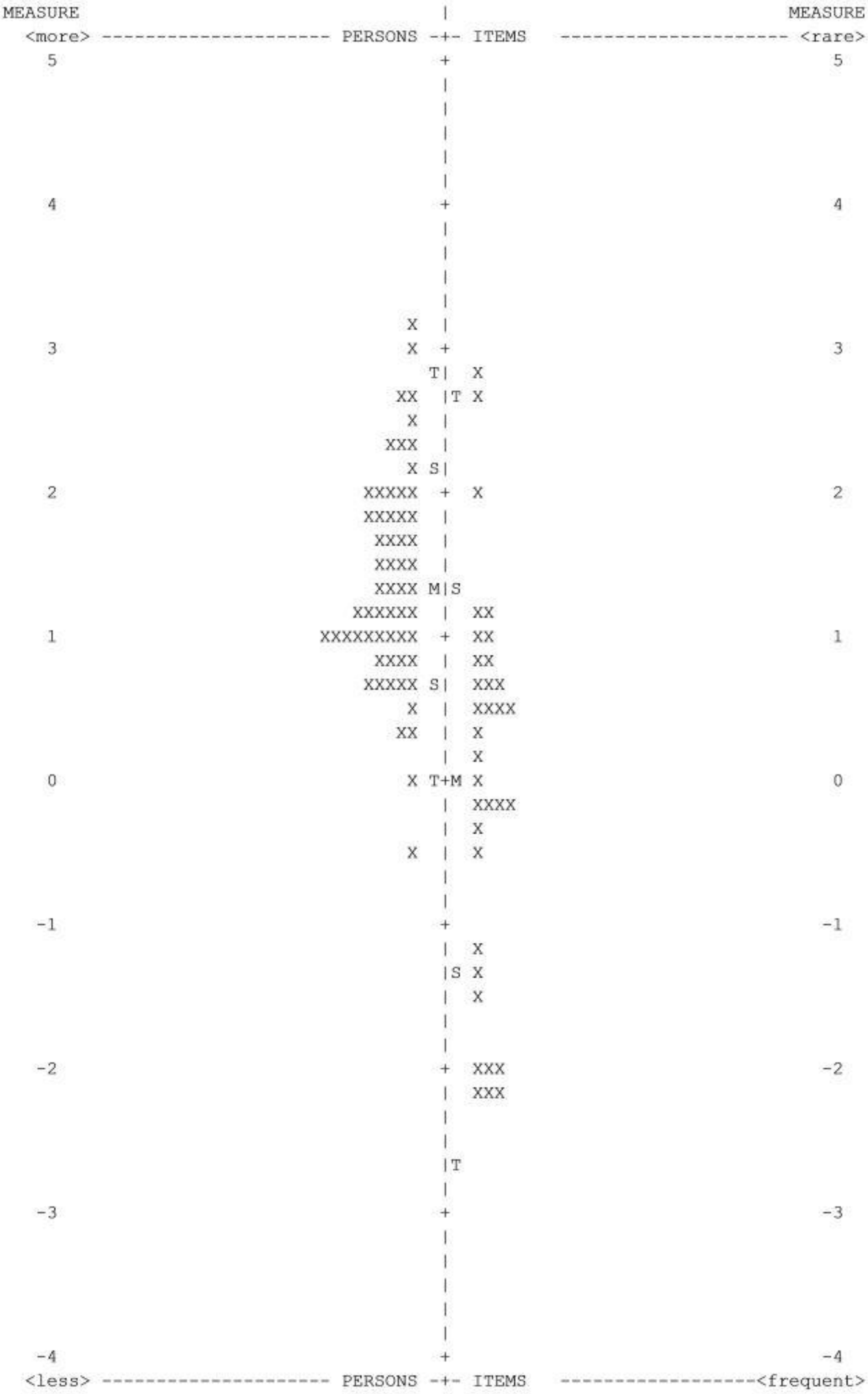
Figure 7. Item Category Measures for Mentor Teacher-Achievement Coaches

```
P      ++---------+---------+---------+---------+---------+---------++
R  1.0 +                                                             +
O      |                                                             |
B      |1                                                            |
A      | 111                                                        4|
B   .8 +    111                                                  444 +
I      |       11                                              444   |
L      |         11                                          44      |
I      |           11                                      44        |
T   .6 +             11                                  44          +
Y      |               11                              44            |
    .5 +                 1              33333333333  4               +
O      |                  11         333           **3               |
F   .4 +                    *22222 33            44   33             +
       |               22222 11  3*222         44       333          |
R      |            222        *3     22      4            33        |
E      |         222         33 11      222 44               333     |
S   .2 +      222          33     11     44*22                  333  +
P      | 22222           33         11 44     222                  33|
O      |2           3333           44*11        2222                 |
N      |       333333         444444     111111      2222222         |
S   .0 +*******444444444444444                 1111111111111*********+
E      ++---------+---------+---------+---------+---------+---------++
       -3        -2        -1         0         1         2         3
                       PERSON [MINUS] ITEM MEASURE
```

Figure 8. Category Probability Curves for Mentor Teacher-Achievement Coaches

Table 6. Item by misfit for mentor teacher-achievement coaches

Entry Number	Infit ZSTD	Outfit ZSTD
33	1.8	3.1
16	2.5	2.8
19	0.5	2.3
10	-2.3	-2.1
12	-2.1	-2.3
20	-2.3	-2.3
3	-2.9	-2.7
5	-3.6	-3.6



```
MEASURE                                 |                               MEASURE
  <more> --------------------- PERSONS -+- ITEMS   --------------------- <rare>
    5                                   +                                   5
                                        |
                                        |
                                        |
                                        |
                                        |
    4                                   +                                   4
                                        |
                                        |
                                        |
                                        |
                                     X  |
    3                                X  +                                   3
                                       T|  X
                                    XX  |T X
                                     X  |
                                   XXX  |
                                     X S|
    2                            XXXXX  +  X                                2
                                 XXXXX  |
                                  XXXX  |
                                  XXXX  |
                                  XXXX M|S
                                XXXXXX  |  XX
    1                        XXXXXXXXX  +  XX                               1
                                  XXXX  |  XX
                                 XXXXX S|  XXX
                                     X  |  XXXX
                                    XX  |  X
                                        |  X
    0                                X T+M X                                0
                                        |  XXXX
                                        |  X
                                     X  |  X
                                        |
                                        |
   -1                                   +                                  -1
                                        |  X
                                        |S X
                                        |  X
                                        |
                                        |
   -2                                   +  XXX                             -2
                                        |  XXX
                                        |
                                        |
                                        |T
                                        |
   -3                                   +                                  -3
                                        |
                                        |
                                        |
                                        |
                                        |
   -4                                   +                                  -4
  <less> --------------------- PERSONS -+- ITEMS   ------------------<frequent>
```

Figure 9. Item-Person Map for Mentor Teacher-Achievement Coaches

5. Conclusions

In analyzing results collected via a survey instrument, it is presumed the respondents have an accurate perception of the construct, rate items according to reproducible criteria, and accurately record their ratings within uniformly spaced levels. In fact, as noted in Wright (1997), ratings are simply responses based on fluctuating personal criteria, the responses are not always interpreted as intended or recorded correctly, and these ratings are ordinal so they do not add up to measures. Rasch analysis produces measures, provides a basis for insight into the quality of the measurement tool and provides information to allow for systematic diagnosis of misfit.

Based on the results of the study, the Differentiated Compensation Survey should be reviewed prior to future data collection. For each of the three groups, the 4-point scale does not function as expected; as most respondents favor categories 3 and 4. Collapsing categories 1 and 2 on the scale would result in probability curves that indicate each category represents a distinct portion of the variable and would likely improve the rating scale diagnostics. For future surveys, the survey construction team might consider a restructuring of the categories creating a dichotomous construction of “Agree” and “Disagree” with a “No Opinion” category to be classified as missing, which as mentioned previously, poses no problems with the Rasch model. Alternatively, the current rating scale of “Agree / Tend to Agree / Tend to Disagree / Disagree” may be reconstituted with category labels exhibiting stronger verbiage so that respondents can more easily distinguish between the categories. In addition to examining the rating scale categories, a few items should be given further attention before being used in subsequent evaluations. In particular, item 33 is highlighted throughout each of the analyses and it is recommended this item be set aside reworked prior to future data collection.

Finally, it is important for the users of the results of the survey to review the hierarchy of items as displayed in the item-person maps (see Figures 3, 6, and 9) to determine if they are meaningful. The items most difficult to endorse (or to agree with) are found at the top end of the figure; the items easiest to agree with are found at the bottom. Inconsistencies between the hierarchies as perceived by the respondents and as expected based on the literature must be examined, with consideration that the source of the problem might stem from either the theoretical basis or the empirical results.

6. Educational Importance

Within the field of education, the development of instruments to assess affective domain constructs has been a problematic area. Surveys are the most common example of self-reported data collection and continue to be one of the most popular research methodologies for graduate studies and published papers in education. Even so, the efficiency and effectiveness of the instrument as a measurement tool is often overlooked or underemphasized.

Bradley and Sampson. (2005) note that the classical test theory model produces a descriptive summary based on statistical analysis, but it is limited if not absent of the capability to assess the quality of the instrument. It is important to begin at the level of measurement and to identify weaknesses that may limit the reliability and validity of the measures made with the instrument. As indicated in the study, Rasch analysis tackles many of the deficiencies of the classical test theory model in that it has the capacity to incorporate missing data, produces validity and reliability measures for person measures and item calibrations, measures persons and items on the same metric, and is person and sample-free.

The survey development team, researchers, organizations, and institutions will benefit from the results of this study as it provides a sound methodology for analyzing rating scale data. The education community will also benefit by receiving better-informed results by collecting data using a more valid and reliable instrument.

7. References

Andrich D., (1978). A Rating Formulation for Ordered Response Categories. *Psychometrika, 43*(4), 561-573.

Becker G., (2001). Controlling Decremental and Inflationary Effects in Reliability Estimation Resulting from Violations of Assumptions. *Psychological Reports, 89*(2), 403-424. doi: 10.2466/pr0.2001.89.2.403

Bond T.G. & Fox C.M., (2001). *Applying the Rasch Model: Fundamental Measurement in the Human Sciences*. Mahwah, NJ: Lawrence Erlbaum.

Bradley K.D. & Sampson S.O., (2005). A case for using Rasch Rating Scale analysis to assess the quality of measurement in survey research. *The Respondent*, 12-13.

Fisher W., (1992). Reliability, Separation, Strata Statistics. *Rasch Measurement Transactions, 6*(3), 238.

Guttman L., (1944). A Basis for Scaling Qualitative Data. *American Sociological Review, 9*(2), 139-150.

Hambleton R.K., Swaminathan, H., & Rogers, H.J., (1991). *Fundamentals of Item Response Theory*. New York: Sage Publications.

Linacre J.M., (2002a). Optimizing Rating Scale Category Effectiveness. *Journal of Applied Measurement, 3*(1), 85-106.

Linacre J.M., (2002b). What do Infit and Outfit, Mean-square and Standardized mean? *Rasch Measurement Transactions, 16*(2), 878.

Linacre J.M., (2004). Winsteps Rasch Measurement computer program (Version 3.51). Beaverton, OR: Winsteps.com.

Rasch G., (1960). *Probabilistic models for some intelligence and attainment tests*. Copenhagen, Denmark.: Danish Institute for Educational Research.

Wright B.D. (1996). Reliability and Separation. *Rasch Measurement Transactions, 9*(4), 472.

Wright B.D., (1997). Fundamental Measurement for Outcome Evaluation. *Physical Medicine and Rehabilitation: State of the Art Reviews, 11*(2), 261-288.

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

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