Challenging the Validity of Higher Education Course Evaluations

Kelly D. Bradley, (Email: kdbrad2@uky.edu), University of Kentucky
James W. Bradley, (Email: will.bradley@kctcs.edu), Bluegrass Community and Technical College

ABSTRACT

In higher education, course evaluations are given much attention, with results directly impacting such events as merit review and tenure/promotion. The accurate presentation and proper use of the evaluation results is a critical issue. The typical course evaluation process involves distributing a Likert-type survey to a class, compiling the data and reporting means/standard deviations (classical test theory approach, CTT). One alternative analytical technique is the Rasch model. A theoretical review of each model and an empirical example utilizing end of semester course evaluations from an introductory statistics course taught at a Midwest community college is presented to demonstrate the step-by-step process of feedback via each model. A contention is made that the CTT summary is not producing a valid picture of the evaluation data. The survey research community and institutions analyzing similar rating scale data will benefit from the results of this study as it provides a sound methodology for analyzing such data. The education community will also benefit by receiving better-informed results.

INTRODUCTION

In higher education, much attention is given to students’ course evaluations. The summarized results of these evaluations often have a direct impact on the faculty teaching the course(s) through processes such as merit review and tenure/promotion. It is not uncommon for higher education administrators to use summarized course evaluation results, usually means and standard deviations, to make curricular decisions and to compare the effectiveness of teaching across the institution. Taking all of this into account, it is critical to consider how course evaluation data are analyzed and consequently what is reported. Hays (1998) writes, “The problem of measurement, and especially of attaining interval scales, is an extremely serious one for the social and behavioral sciences. It is unfortunate that in their search for quantitative methods researchers sometimes overlook the question of level of measurement and tend to read quite unjustified meaning in to their results” (p. 71). When researchers develop a group of items intended to assess a construct, administer the items to a nonrandom sample of respondents, and sum the ratings, certain assumptions are put in place:

- Each item contributes equally to the measure of that construct, implying all items are of equal importance.
- Each item is measured on the same interval scale.
- 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 often problematic, in survey research methods (Bond & Fox, 2001, Sampson & Bradley, 2003). For example, in practice the scale is actually ordinal, so categories are not necessarily spaced equally.

When an instructor receives a descriptive summary of the students’ course evaluations is this an adequate picture of the reported information? Here a contention is made that the answer is ‘No’. The instructor, and others utilizing the information, is only receiving a small piece of information, and the information is limited in scope as it presents only student perceptions. It seems that those using the results would be interested in measures on the items themselves, as well as an assessment of the actual measurement instrument. Thus, this paper takes the data produced from a set of student evaluations and analyzes them via a classical test theory (CTT) and a Rasch theory approach in an attempt to illustrate that a traditional summarized report, based on means and standard deviations, is not presenting the 'whole picture', and in some cases, may not be providing valid information.

THEORETICAL FRAMEWORK

Rasch versus Classical Test Theory (CTT) Approach

Researchers often utilize the classical test theory model in analyzing the rating scale data produced via the selected-response survey. As noted in Smith (2000), the classical test theory model, sometimes called the true score model, requires complete records to make comparisons of items on the evaluation instrument. Even if this is attained, the issue of sample-dependence between estimates of an item's difficulty to endorse and a respondent's willingness to endorse surface. This is problematic since it makes the estimates for the items dependent on the rater-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. The CTT approach produces a single standard error of measurement for the composite of the ratings, making it inadequate and potentially misleading.

The Rasch model, introduced by Georg Rasch (1960), provides estimates for persons and items that are freed from the sampling distribution of the sample employed [given the data fit the model], meaning there is no dependence on the particulars of the evaluation or of the sample being measured. Rasch measurement 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. Respondents and items are measured on the same metric, allowing 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 in that the probability of providing a certain response defines an order of respondents and items (Smith, E., 2000; Wright, 1997; 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. The model provides one mathematically sound alternative to analyzing traditional classroom evaluations.

METHODS

The data utilized in the examples presented were collected during the spring 2000 Quarter at a Midwest community college. Nineteen students enrolled in an Introductory Statistics course were asked to fill out a course evaluation (See Appendix), which consisted of 28 questions. Each question was written in the form of a statement, such as 'The instructor clearly defined course requirements'. The students were asked to rate their agreement with each statement using on the following Likert-type scale: 1 = strongly disagree, 2 = disagree, 3 = neutral, 4 = agree, and 5 = strongly agree. Students were also provided space on the back of the evaluation to enter comments. Those comments are not used in this analysis.

Before proceeding, it is important to gain a general understanding of the data (see Appendix). Missing data was recorded as "*" and treated as missing. Given the evaluation is a collection of perceptions; it is reasonable to believe that a student may not have an opinion on every item. Thus, means or other values were not imputed. Valid responses were coded as 1, 2, 3, 4 or 5, as described above. The assignment of numbers is simply a form of ranking, where 2 is "more" than 1; however, it is unclear how much more. Given this measurement concept, the data should be considered ordinal. Even if a mathematical transformation is performed on the data, the results cannot necessarily be interpreted as a statement about the true magnitude of the response (Hays, 1988).

Research Objective

In this paper data are analyzed using two approaches, CTT and Rasch. It is our contention that employing the Rasch model is essential in addressing the many weaknesses of the CTT approach. First, scores obtained from the same set of items require complete records in order to be compared in the CTT setting. Rasch measurement has the ability to incorporate missing data. Next, there is only a single standard error of measurement for the scores in the CTT setting, where in the Rasch setting we see measures for both person measures and item calibrations. A major concern with producing a statistical summary via CTT is that the raw scores for persons and items and linear transformations are not on a linear interval scale, which violates the underlying assumption of the model. Rasch measurement provides estimates for person and items that are freed from the sampling distribution. Using the Rasch model allows for the prediction of the outcome of the interaction between a given person and a given item. This cannot be done in the CTT realm, since different metrics are used for person and items. In addition, there are few techniques in CTT for validating response patterns. Within the Rasch analysis, results for each statement and person are provided in order to investigate patterns within the responses. Finally, Rasch measurement makes it possible to identify the optimal number of points for rating scales (Smith, 2000). All of this occurs within the framework of the data fitting the Rasch model.

The CTT Report

The Midwest community college includes the following in a report that they produce: total valid responses, average rating, calculated by producing means for each item, frequency counts for each rating and the corresponding percentage. In addition, overall means were computed. A sample of the report is presented below in Table 1.

Table 1: Faculty Evaluation Report

Question	Valid Responses	Average Rating	#1	#2	#3	#4	#5
Clear Requirements	19	4.4	0	2	1	4	12
Clear Grading System	19	4.4	0	1	2	5	11
Thorough Knowledge	18	4.8	0	0	0	4	14
Enthusiastic	19	4.6	0	0	1	6	12
Examples Used	19	4.6	0	0	1	6	12
Major Points Emphasized	19	4.5	0	0	1	8	10
Material Explained	19	4.1	1	0	2	9	7
Encouraged Questions	19	4.5	0	0	2	6	11
Answered Questions	19	4.6	0	1	0	4	14
Helped Those in Need	19	4.6	0	0	1	6	12
Respect for Students	19	4.7	0	0	0	5	14
Maintained Atmosphere	19	4.6	0	0	1	5	13
Used Time Well	19	4.6	0	0	0	7	12
Available Class Period	19	4.6	0	0	1	5	13
Retention of Work (1 Week)	14	4.5	0	1	0	6	10
Recommend Instructor	17	4.5	1	0	0	5	11
Text Helped	19	4.2	0	0	4	8	7
Classroom Adequate	19	4.3	1	0	1	7	10
Teacher-Student Discussion	19	4.3	0	0	5	3	11
Informed of Progress	19	3.9	1	1	5	3	9
Challenged to Think	19	4.5	0	0	1	5	12
Reasonable Tests	19	4.5	0	0	2	6	11
Class Preparation	19	4.6	0	0	1	5	13
High Standards Held	19	4.5	0	0	1	7	11
Complete Homework	19	4.6	0	0	2	4	13
Attend Class	19	4.7	0	1	0	2	16
Improve Problem-Solving	19	4.2	0	1	3	7	8
Appropriate Level	18	4.4	0	0	2	7	9

Valid responses range from 17 to 19 persons per item. The mean rating for each item ranges from 3.9 to 4.8. Additionally, students responded in highest frequency with ratings of 4 and 5. An overall mean rating of 4.5 was calculated for this instructor, indicating high satisfaction and a solid teaching performance as interpreted by the Community College.

It becomes evident that very little information is provided. The instructor is told that their average rating is a 4.5, which is approaching the highest rating of 5. Still, a rating of 4.5 is not described on the original rating scale, so it is unclear what this means exactly. Furthermore, nothing is reported about 'student harshness'. Did some students consistently rate the instructor low while others rated him high? It is the case that the institution could have produced more statistics within the true score model, including standard deviations and a standard error of measurement. These would have provided more insight, but only on a single dimension. Turning our attention to the Rasch model, we can investigate the instrument, persons, and items.

Employing Rasch Model

A one-parameter Item Response Theory model was utilized, commonly known as the Rasch Model, using Winsteps software (Linacre, 2004 version 3.51). Winsteps implements 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 & Masters, 1982). The Rasch model used in Winsteps for this analysis is the polytomous "Rating Scale" model with the equation: $\log\left(P_{nij} \middle/ P_{ni\left(j-1\right)} \right) = B_n - D_i - F_j$, where P_{nij} is the probability that person n encountering item i is observed in category j, B_n is the "ability" or rater-severity measure of person n, D_i is the difficulty-to-endorse measure of item i, and F_j is the "calibration" measure of category j relative to category $P_{nij}\, B_n\, D_i\, F_j \left(j-1\right)$ (Linacre, 2004).

The Rasch model uses the sum of the item ratings simply as a starting point for estimating probabilities of those responding. 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 (Linacre, 1999). 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 (Linacre, 2002). 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 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, p. 29). Bond and Fox explain that employing Rasch techniques allows for the ordering of respondents along this continuum of willingness to endorse items and orders items along a continuum according to their difficulty to endorse. "Based on this logic of order, the Rasch analysis software programs perform a logarithmic transformation on the item and person data to convert the ordinal data to yield interval data…actual item and person performance probabilities determine the interval sizes" (p. 29).

The use of any Rasch measurement model specifies two requirements: (1) Most of the items must provoke data along the same underlying construct. Here, the instrument claims to measure 'the quality or effectiveness of instruction' via the performance of the instructor. (2) The probability of responding correctly to one item must not be influenced by the particular response to another item (Wright, 1996). Since there are not necessarily 'correct' responses on an evaluation this is not of grave concern. However, with perception type of instruments, it is always important to consider any participant that may complete the evaluation from a socially desirable standpoint – meaning that an individual responds in a manner they feel would either match with their peers or be approved of by the instructor.

The Rasch analysis of the Community College student evaluations utilizes the same 5 category Likert-type rating scale as presented above in the CTT report, ranging from 1 = strongly disagree to 5 = strongly agree. The data set is comprised of 28 items and 19 persons. This analysis uses the WINSTEPS software (Wright and Linacre, 2000 version 3.02) and is based on the two-facet rating scale model where the parameters estimated are 19 person measures, 28 item measures and 4 category thresholds relating to the transition points between the 5 response categories (See Appendix for Sample Program Code).

RESULTS/DISCUSSION

Table 3.1 (See Figure 1) is a good place to begin interpreting the results of the Rasch analysis as it provides an overview of the reliability estimates. The real separation reliability is highlighted below and is comparable to a

Cronbach's alpha estimate. Here, 'real' indicates that the estimated standard errors of measurement have been adjusted for any misfit encountered in the data. The real person reliability of 0.85 suggests that the scale discriminates well between the persons. The real item separation reliability of 0.34 suggests that the items may not be creating a well-defined variable. INFIT and OUTFIT ZSTD statistics are also reported in Table 3.1. OUTFIT ZSTDs are the standardized unweighted item and person fit statistics. These estimates are sensitive to unexpected rare extremes. INFIT ZSTDs are standardized information-weighted item and person fit statistics. These estimates are sensitive to irregular inlying patterns. When the data fit the model, these statistics are approximately t-statistics. For this setting, the approximate t-statistics would have a mean of 0 and standard deviation of 1. Here (highlighted), the mean is close to 0 in both cases; however the standard deviation is high suggesting that there are some items that misfit and there is more variability in the fit of the students than expected (Wright and Masters, 1982).

Figure 1: WINSTEPS Table 3.1

```
                       TABLE 3.1 Teacher Eval Run
INPUT: 19 PERSONS, 28  ITEMS  ANALYZED: 17 PERSONS, 28  ITEMS, 5 CATS      v3.02
--------------------------------------------------------------------------------

                SUMMARY OF     17 MEASURED (NON-EXTREME) PERSONS
+-------------------------------------------------------------------------------+
|           RAW                          MODEL         INFIT        OUTFIT      |
|          SCORE     COUNT     MEASURE   ERROR      MNSQ   ZSTD   MNSQ   ZSTD  |
|-------------------------------------------------------------------------------|
| MEAN     121.9       27.6        2.81     .47      1.10    -.1   1.08    -.1  |
| S.D.      14.4         .6        1.80     .28       .74    2.1    .82    2.2  |
| MAX.     139.0       28.0        5.73    1.02      3.09    4.9   3.43    4.9  |
| MIN.      89.0       26.0         .10     .22       .24   -3.6    .26   -3.5  |
|-------------------------------------------------------------------------------|
| REAL RMSE    .58  ADJ.SD    1.70  SEPARATION  2.92  PERSON RELIABILITY  .90  |
|MODEL RMSE    .55  ADJ.SD    1.71  SEPARATION  3.13  PERSON RELIABILITY  .91  |
| S.E. OF PERSON MEAN = .45                                                     |
| WITH 2 EXTREME = 19 PERSONS  MEAN = 3.25,  S.D. = 2.12                        |
| REAL RMSE    .81  ADJ.SD    1.96  SEPARATION  2.43  PERSON RELIABILITY  .85  |
|MODEL RMSE    .79  ADJ.SD    1.97  SEPARATION  2.51  PERSON RELIABILITY  .86  |
+-------------------------------------------------------------------------------+
                     MAXIMUM EXTREME SCORE:      2 PERSONS
                            VALID RESPONSES:  98.7%

                       SUMMARY OF     28 MEASURED  ITEMS
+-------------------------------------------------------------------------------+
|           RAW                          MODEL         INFIT        OUTFIT      |
|          SCORE     COUNT     MEASURE   ERROR      MNSQ   ZSTD   MNSQ   ZSTD  |
|-------------------------------------------------------------------------------|
| MEAN      74.0       16.8         .00     .45      1.01    -.4   1.09    -.2  |
| S.D.       4.3         .6         .65     .07       .87    1.5   1.07    1.2  |
| MAX.      80.0       17.0        1.45     .61      3.72    3.3   4.05    2.2  |
| MIN.      65.0       15.0       -1.14     .33       .28   -2.2    .25   -1.5  |
|-------------------------------------------------------------------------------|
| REAL RMSE    .53  ADJ.SD     .38  SEPARATION   .71   ITEM  RELIABILITY  .34  |
|MODEL RMSE    .45  ADJ.SD     .46  SEPARATION  1.01   ITEM  RELIABILITY  .50  |
| S.E. OF  ITEM MEAN = .12                                                      |
+-------------------------------------------------------------------------------+
```

Seeing that the misfit of items surfaces as a concern, attention is given to Table 14.1 (See Figure 2). The table presents a summary of the individual item statistics. Values less than –2 are considered to be 'muted', meaning redundancy or error trends exist. Values greater than 2 are considered to be 'noisy', an indication of unexpected or inconsistent irregularities (Linacre, 2000). The statistics reveal there are five items falling above or below this cutoff (highlighted below), which warrant further review.

Figure 2: WINSTEPS Table 14.1

```
                         TABLE 14.1 Teacher Eval Run
INPUT: 19 PERSONS, 28  ITEMS  ANALYZED: 17 PERSONS, 28  ITEMS, 5 CATS      v3.02
--------------------------------------------------------------------------------
                        ITEMS STATISTICS:  ENTRY ORDER
       +------------------------------------------------------------------+
       |ENTRY    RAW                               |   INFIT  |  OUTFIT  |SCORE|
       |NUMBER  SCORE  COUNT  MEASURE      ERROR|MNSQ  ZSTD|MNSQ  ZSTD|CORR.|
       |------------------------------------+----------+----------+-----+
       |      1     73     17      .43        .40|1.81   1.5|1.74   1.1|  .53|
       |      2     73     17      .43        .40|1.07    .2|2.09   1.4|  .56|
       |      3     76     16    -1.14        .61| .69   -.7| .56   -.4|  .54|
       |      4     77     17     -.32        .47| .46  -1.4| .38  -1.0|  .75|
       |      5     77     17     -.32        .47|1.08    .2|1.31    .4|  .52|
       |      6     75     17      .09        .43| .28  -2.2| .28  -1.5|  .82|
       |      7     67     17     1.23        .34| .71   -.8| .57  -1.1|  .76|
       |      8     76     17     -.10        .45| .92   -.2| .93   -.1|  .56|
       |      9     78     17     -.55        .50|1.10    .2| .60   -.5|  .65|
       |     10     77     17     -.32        .47| .54  -1.1| .61   -.6|  .70|
       |     11     80     17    -1.12        .57| .43  -1.5| .29   -.9|  .69|
       |     12     78     17     -.55        .50| .36  -1.8| .26  -1.2|  .77|
       |     13     78     17     -.55        .50| .28  -2.1| .25  -1.2|  .79|
       |     14     78     17     -.55        .50| .45  -1.5| .31  -1.1|  .75|
       |     15     66     15      .08        .46|2.42   2.0|3.71   2.2|  .26|
       |     16     66     15      .22        .45|3.72   3.3|3.45   2.2|  .24|
       |     17     71     17      .73        .37|1.04    .1|1.98   1.5|  .49|
       |     18     73     17      .43        .40|1.07    .2|1.11    .2|  .62|
       |     19     72     17      .59        .38| .57  -1.2| .60   -.8|  .78|
       |     20     65     17     1.45        .33| .74   -.8| .87   -.3|  .78|
       |     21     75     17      .09        .43| .66   -.8| .44  -1.1|  .75|
       |     22     75     17      .09        .43| .42  -1.6| .36  -1.3|  .79|
       |     23     78     17     -.55        .50| .58  -1.0| .41   -.9|  .72|
       |     24     76     17     -.10        .45| .31  -2.0| .29  -1.4|  .80|
       |     25     77     17     -.32        .47|2.20   1.9|1.98   1.0|  .28|
       |     26     79     17     -.82        .53|3.23   3.0|4.05   1.8|  .13|
       |     27     68     17     1.11        .35| .51  -1.5| .53  -1.2|  .81|
       |     28     69     16      .34        .42| .53  -1.2| .45  -1.1|  .77|
       |------------------------------------+----------+----------+-----+
       | MEAN      74.    17.      .00        .45|1.01   -.4|1.09   -.2|     |
       | S.D.       4.     1.      .65        .07| .87   1.5|1.07   1.2|     |
       +------------------------------------------------------------------+
```

As highlighted above, the five items are:

- Item (6) The instructor emphasized major points.
- Item (13) The instructor used class time well.
- Item (15) Tests or assignments were returned within a week.
- Item (16) I would recommend this instructor to someone wanting to learn.
- Item (26) I regularly attended class.

When reviewing the items, it could be argued that item 26 is not tapping into the instructor's effectiveness. Instead it seems to be more of a demographic variable related to the student completing the survey. This alone could constitute the misfit. Reflecting back to the CTT report, reporting a mean across all items does not seem relevant, as an item like 26 illustrates. This item is not clearly an evaluation indicator of the instructor.

Table 9.1 (See Figure 3) provides a visual display of the OUTFIT information presented in Figure 2. The largest misfits, A, B, and C, are plotted at the top of the graph. Viewing the display, and considering that the largest misfits are shown at the bottom of the figure, there are not large overfits. The person distribution is shown at the bottom of the graph, with the mean of the person distribution being marked with a vertical line. Here, the mean is close to 3. Furthermore, most items do not even fall within one standard deviation, represented by S on the horizontal axis. Most of the items are falling away from the mean, indicating the items are extreme in ‘difficulty’ for the students (Linacre, 2000). Likely, respondents are having difficulty in understanding the meaning of these items.

Figure 3: WINSTEPS Table 9.1

```
                           TABLE 9.1 Teacher Eval Run
INPUT: 19 PERSONS, 28  ITEMS  ANALYZED: 17 PERSONS, 28  ITEMS, 5 CATS      v3.02
--------------------------------------------------------------------------------
        -2      -1       0       1       2       3       4       5       6
        ++-------+-------+-------+-------+-------+-------+-------+-------++
      5 +                                       |                         +   5
        |                                       |                         |
        |                                       |                         |
  I     |                                       |                         |
  T   4 +         A                             |                         +   4
  E     |                 C                     |                         |
  M     |                  B                    |                         |
        |                                       |                         |
  O   3 +                                       |                         +   3
  U     |                                       |                         |
  T     |                                       |                         |
  F     |                                       |                         |
  I   2 +-------------D-----E--F----------------|-------------------------+   2
  T     |                   G                   |                         |
        |                                       |                         |
  M     |             H                         |                         |
  N   1 +---------------K---I-------------------|-------------------------+   1
  S     |                            L          |                         |
  Q     |       N   kJmh  n j l   iM            |                         |
        |       f   agd c eb                    |                         |
      0 +                                       |                         +   0
        ++-------+-------+-------+-------+-------+-------+-------+-------++
        -2      -1       0       1       2       3       4       5       6
                                    ITEM MEASURE

  PERSON                  1  1   2 1  1    1 111  2    1         1     3 2
                   T             S              M             S
```

Table 14.3 (See Figure 4) provides frequency counts for each item by distracter (Wright and Masters, 1982). For example looking at item 6, there was 1 neutral (3), 8 agree (4), and 10 strongly agree (5). Considering the items of ‘concern’, besides item 6, the response pattern is non-consistent, having no clear distribution. Two items also contain omits.

Table 10.4 (Figure 5) is helpful in diagnosing the misfit. It contains a listing of the most unexpected responses for the most misfitting items. Responses producing large residuals are shown with the actual response; whereas, expected responses are indicated with ‘.’. Persons with a willingness to agree with the statements (high scores) are shown on the left, while persons with a willingness to disagree (low scores) are shown on the right. Considering this information, that student ‘10’, a student likely to agree, was responsible for two of the unexpected responses (highlighted below).

Part of the misfit could be attributed to the actual rating scale and how the students perceive it or apply it. Table 21.1 (See Figure 6), allows one to analyze the fit of the steps. Specifically, it provides an answer to the question: is the distance between a rating of 1 and 2, the same as the distance between 2 and 3? This is essentially a ‘test’ of the equidistant assumption in CTT. Looking at the plot, a smooth transition does not exist between category 2, disagree, and category 3. These probability curves suggest that the students completing the evaluations are using only three, possibly four (if 2 and 3 were combined), of the five categories offered (illustrated by the curves for the 1’s, 4’s, and 5’s). This set of curves is applied to all items with an adjustment for item difficulty to position each item’s probability curves on the logit metric (Linacre, 2000). In the CTT approach, there is no consideration of this idea, one that is of critical importance in collection of perceptions via Likert-type scales.

Figure 4: Items Of 'Concern' From Table 14.3

```
                      TABLE 14.3 Teacher Eval Run
INPUT: 19 PERSONS, 28  ITEMS  ANALYZED: 17 PERSONS, 28  ITEMS, 5 CATS      v3.02
--------------------------------------------------------------------------------
               ITEMS OPTION/DISTRACTOR FREQUENCIES:  ENTRY ORDER
            +---------------------------------------------------------+
            |ENTRY   DATA  SCORE |  DATA       |    USED      AVERAGE |
            |NUMBER  CODE  VALUE |  COUNT    % |   COUNT   %  MEASURE |
            |--------------------+-------------+----------------------+
            |    6   3         3 |      1    5 |      1   6       .10 |
            |        4         4 |      8   42 |      8  47      1.61 |
            |        5         5 |     10   53 |      8  47      4.36 |
            |                                                         |
            |   13   4         4 |      7   37 |      7  41      1.11 |
            |        5         5 |     12   63 |     10  59      4.01 |
            |                                                         |
            |   15   2         2 |      1    6 |      1   7      2.27 |
            |        4         4 |      6   35 |      6  40      2.14 |
            |        5         5 |     10   59 |      8  53      3.45 |
            |        MISSING *** |      2   10 |      2  12      2.58 |
            |                                                         |
            |   16   1         1 |      1    6 |      1   7      2.68 |
            |        4         4 |      5   29 |      5  33      1.84 |
            |        5         5 |     11   65 |      9  60      3.60 |
            |        MISSING *** |      2   10 |      2  12      1.76 |
            |                                                         |
            |   26   2         2 |      1    5 |      1   6      3.10 |
            |        4         4 |      3   16 |      3  18      1.48 |
            |        5         5 |     15   79 |     13  76      3.10 |
            +---------------------------------------------------------+
```

Figure 5: WINSTEPS Table 10.4

```
                      TABLE 10.4 Teacher Eval Run
INPUT: 19 PERSONS, 28  ITEMS  ANALYZED: 17 PERSONS, 28  ITEMS, 5 CATS      v3.02
--------------------------------------------------------------------------------

   MOST MISFITTING RESPONSE STRINGS
   ITEM                                           OUTMNSQ  |PERSON
                                                            |11   1 1 1
                                                            |914201398358
                                                        high------------
       26 Regulary attended class                     4.05 A|......2.....
       16 Would recommend this instructor             3.45 B|........1 ..
       15 Test/assignments returned within a week     3.71 C|.4......  2.
       25 Regulary completed homework                 1.98 D|......3...3.
        2 Defined Grading System                      2.09 E|4........3..
       17 Text helped me learn                        1.98 F|..44........
        1 Defined Course Requirements                 1.74 G|........2...
        5 Examples to help understand                 1.31 H|.....3......
       18 Adequate Classroom facilities               1.11 I|...4.......1
        9 Answered questions related to subject        .60 J|...........2
        8 Encouraged students to ask questions         .93 K|....4.......
        3 Knowledge of subject                         .56 N|.......4...
       10 Helped those who needed help                 .61 m|....4.......
       27 Helped me improve my problem solving skills  .53 i|......3.....
                                                            |--------low-
                                                            |114211191358
                                                            |91   0 3 8
```

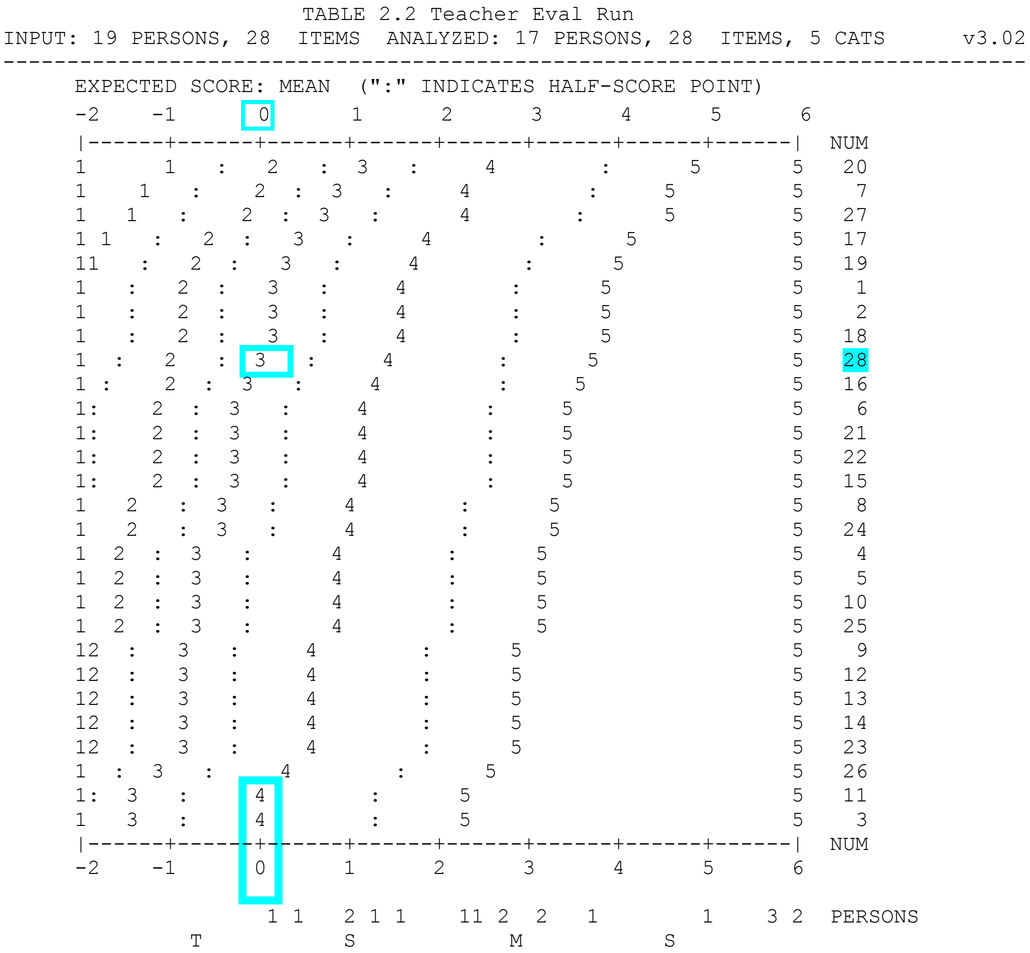
Figure 6: WINSTEPS Table 21.1

```
                           TABLE 21.1 Teacher Eval Run
INPUT: 19 PERSONS, 28  ITEMS  ANALYZED: 17 PERSONS, 28  ITEMS, 5 CATS      v3.02
--------------------------------------------------------------------------------
         CATEGORY PROBABILITIES: MODES - Step measures at intersections
       P      ++-------+-------+-------+-------+-------+-------+-------++
       R  1.0 +                                                         +
       O      |                                                         |
       B      |                                                         |
       A      |111                                                    55|
       B   .8 +   11                                               555  +
       I      |     11                                           55     |
       L      |       1                                         5       |
       I      |        11                                     55        |
       T   .6 +          1                  444444444       55          +
       Y      |           1               44         444   5            |
         .5 +           1            44              4*5             +
       O      |             1          4                5 4             |
       F   .4 +              1        4               55   44           +
          |               1  3333*33             5       44         |
       R      |            2222*3  44   333        55          44       |
       E      |        2222  332*2*        33    55              44     |
       S   .2 +     222    33    * 22        3355                  444  +
       P      |  222     33    44 11 22      55333                    44|
       O      |22      33    44     11 222555     3333                  |
       N      |   33333  4444        5****22222       333333            |
       S   .0 +**********555555555555     11111*************************+
       E      ++-------+-------+-------+-------+-------+-------+-------++
           -3      -2      -1       0       1       2       3       4
        PERSON [MINUS]  ITEM MEASURE
```

The probability curves above are used to form a type of calibration tool. Table 2.2 (Figure 7) can be used to produce a quick estimate of an expected score for a person at any measure by finding the person's measure on the horizontal axis and drawing a perpendicular line through that point. The response categories nearest that vertical line are the person's most likely response. Here a person with a measure of 0.0 (marked by a box) would have an expected score of 4 (agree) on the two items most difficult to endorse (items 3 and 11). The same person, with a measure of 0.0, would be expected to respond to question 28 (highlighted) with a neutral rating (3). This technique may be used when dealing with missing data, allowing institutions to make more accurate extrapolations from the existing data.

Finally, Table 22.1 (See Figure 8) provides a Guttman scalogram of the raw data. The items are ordered from hardest to endorse to easiest to endorse across the columns. The persons, here students completing the evaluation, are ordered from highest raw score (most likely to endorse) to lowest raw score (least likely to endorse) down the rows. Three of the misfitting items, 3, 26 and 13 (highlighted below) can be identified as some of the more difficult items to endorse (Linacre, 2000). This is a distinct advantage of this model as compared to the CTT model. One criticism of higher education evaluations is the lack of a good comparison between classes and instructors. The Rasch model provides this opportunity and would allow institutions to adjust for easy or difficult student raters, making the comparison between classes and instructors more accurate and fair.

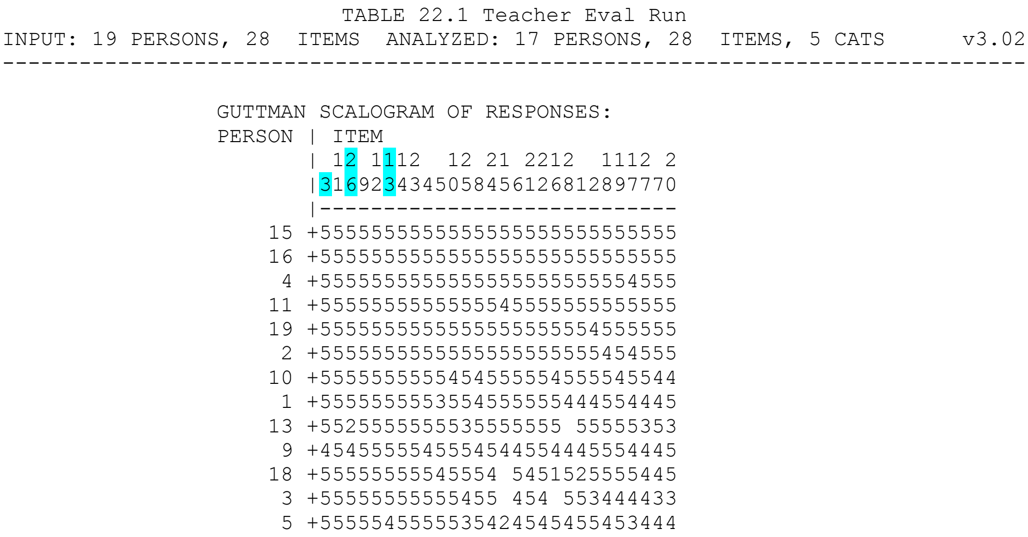
Figure 7: WINSTEPS Table 2.2



```
                       TABLE 2.2 Teacher Eval Run
INPUT: 19 PERSONS, 28  ITEMS  ANALYZED: 17 PERSONS, 28  ITEMS, 5 CATS      v3.02
--------------------------------------------------------------------------------
      EXPECTED SCORE: MEAN  (":" INDICATES HALF-SCORE POINT)
      -2    -1      0      1      2      3      4      5      6
      |------+------+------+------+------+------+------+------|  NUM
      1      1   :   2   :  3   :     4        :      5       5   20
      1    1   :    2  :  3   :     4         :     5         5    7
      1   1   :    2  :  3   :      4        :      5         5   27
      1 1   :   2  :   3   :     4        :      5            5   17
      11   :   2  :   3   :     4        :      5             5   19
      1   :   2  :   3   :     4        :      5              5    1
      1   :   2  :   3   :     4        :      5              5    2
      1   :   2  :   3   :     4        :      5              5   18
      1  :   2   :  3   :     4        :      5               5   28
      1 :    2  :  3   :     4         :     5                5   16
      1:    2  :  3   :     4         :     5                 5    6
      1:    2  :  3   :     4         :     5                 5   21
      1:    2  :  3   :     4         :     5                 5   22
      1:    2  :  3   :     4         :     5                 5   15
      1   2   :  3   :     4        :      5                  5    8
      1   2   :  3   :     4        :      5                  5   24
      1  2  :  3   :      4        :      5                   5    4
      1  2  :  3   :      4        :      5                   5    5
      1  2  :  3   :      4        :      5                   5   10
      1  2  :  3   :      4        :      5                   5   25
      12  :   3   :     4        :      5                     5    9
      12  :   3   :     4        :      5                     5   12
      12  :   3   :     4        :      5                     5   13
      12  :   3   :     4        :      5                     5   14
      12  :   3   :     4        :      5                     5   23
      1  :  3   :     4        :      5                       5   26
      1:  3   :     4        :      5                         5   11
      1   3   :     4        :      5                         5    3
      |------+------+------+------+------+------+------+------|  NUM
      -2    -1      0      1      2      3      4      5      6

                     1 1   2 1 1    11 2  2   1        1    3 2  PERSONS
               T           S            M           S
```

Figure 8. WINSTEPS Table 22.1



```
                          TABLE 22.1 Teacher Eval Run
INPUT: 19 PERSONS, 28  ITEMS  ANALYZED: 17 PERSONS, 28  ITEMS, 5 CATS      v3.02
--------------------------------------------------------------------------------

                 GUTTMAN SCALOGRAM OF RESPONSES:
                 PERSON | ITEM
                        | 12 1112  12 21 2212  1112 2
                        |3169234345058456126812897770
                        |----------------------------
                     15 +5555555555555555555555555555
                     16 +5555555555555555555555555555
                      4 +5555555555555555555555554555
                     11 +5555555555555545555555555555
                     19 +5555555555555555555554555555
                      2 +5555555555555555555555454555
                     10 +5555555555454555554555545544
                      1 +5555555553554555555444554445
                     13 +5525555555355555555 55555353
                      9 +4545555545554544554445554445
                     18 +55555555545554 5451525555445
                      3 +55555555555455 454 553444433
                      5 +5555545555535424545455453444
```

```
14 +5454444445444444445455444443
 7 +4445444444444544444444434444
 6 +4454444444445444444444433433
12 +555444434445345443 354434333
17 +4444443434344444344333434242
 8 + 452344444454353235422134311
   |----------------------------
   |3129111245128216221212111272
   | 16 2343  05 45 1268  8977 0
```

CONCLUSIONS

The WINSTEPS output contains many other tables that have been omitted from this discussion. Even with the selected overview of information presented, it becomes clear that this type of analysis provides more accurate, fair and useful results in analyzing instructor evaluations that employ a Likert-type scale. Institutions typically collect this information via paper-and-pencil survey instruments that attempt to measure ‘the quality or effectiveness of instruction’. Once collected, the common approach is to produce means and standard deviations and then make comparisons across classes and instructors. As illustrated through the Rasch example presented above, there are many weaknesses of this approach. Using the two-facet Rasch model approach, provides a detailed review of the instrument’s rating scale, as well as an accurate description of person measures and item measures.

In this example, Rasch results indicate students are utilizing only three to four categories, even though there are five in total. It was also discovered that student 10 was responding in an unexpected pattern. Furthermore, items were found to be misfitting within the instrument, either not measuring as intended or too difficult for responding students to endorse. It could be argued that the items are not forming a well-defined variable, indicating that this institution [and likely most institutions] needs to form a more defined construct of ‘instructor quality or effectiveness’.

In analyzing results collected via instructor evaluations (here and in most surveys), 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. One mathematical alternative to the commonly CTT approach of reporting means and standard deviations for instructor evaluation items is the Rasch model. Given the importance that many institutions are giving to instructor evaluation results, it becomes of critical importance to make accurate and fair conclusions based upon the data. Data-driven decision making is only reliable and valid when proper analysis is conducted. Here, a contention is made that the Rasch approach, in the family of Item Response Theory models, provides this precision.

EDUCATIONAL IMPORTANCE

Within the field of education, the development of instruments to assess affective domain constructs has been a problematic area (Aiken, 1996; Martin, 1983). The usefulness, more specifically proper use, of evaluation instruments is often overlooked or underemphasized. The typical course of action is to distribute a basic Liker-type survey to a classroom of college students, collect the data and report means and standard deviations [often without even controlling for other influential variables]. As noted by Sampson and Bradley (2003) and Bradley and Sampson (2005), the CTT 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 CTT model.

The survey research community and institutions analyzing similar rating scale data will benefit from the results of this study as it provides a sound methodology for analyzing such data. The education community will also benefit by receiving better-informed results. Simply stated, our argument is this. Descriptive CTT statistics are not portraying an accurate picture of rating scale data. If institutions are going to continue the routine of evaluating an instructor using a rating scale format, then the data should be analyzed in such a way to give complete and accurate feedback, assuring reliable and valid results. CTT provides a single snapshot, while the Rasch approach provides the complete album.

REFERENCES

1. Bond, T and Fox, C. (2001). *Applying the Rasch model: Fundamental measurement in the human sciences*. Mahwah, NJ: Lawrence Erlbaum Associates.
2. Bradley, K. D. and Sampson, S. (2005). A case for using a Rasch model to assess the quality of measurement in survey research. *The Respondent*, 12-13.
3. Hays, W. L. (1988). *Statistics* (4th ed.). Fort Worth: Holt Rinehart and Winston.
4. Linacre, J. (1999). *A User's Guide to Facets Rasch Measurement Computer Program*. Chicago, IL: MESA Press.
5. Linacre, J.M. (2000). Handout from Rasch Measurement Training Seminar. Chicago.
6. Linacre, J.M. and Wright, B.D. (2000). *A user's guide to WINSTEPS: Rasch Model Computer Program*. Chicago: MESA Press.
7. Linacre, J. (2002). Facets, factors, elements and levels [Electronic version]. *Rasch Measurement Transactions, 16* (2), 880.
8. Rasch, G. (1980). *Probabilistic models for some intelligence and attainment tests.* Chicago: The University of Chicago Press (original work published in 1960).
9. Sampson, S. & Bradley, K. D. (November, 2003). Rasch analysis of educator supply and demand rating scale data [Electronic Version]. *Research Methods Forum.* Available at: http://aom.pace.edu/rmd/2003forum.html
10. Smith, E., Jr. (2000). *Rasch Measurement Models.* Paper presented at An Introduction to Rasch Measurement: Theory and Applications, Chicago.
11. Wright, B.D. (1996). Comparing Rasch measurement and factor analysis. *Structural Equation Modeling*, 3, 3 – 24.
12. Wright, B.D. and Linacre, J.M. (2000). WINSTEPS, version 3.02.
13. Wright, B. D., and Masters, G. N. (1982). *Rating scale analysis*. Chicago, IL: MESA Press.
14. Wright, B. (1997). Fundamental measurement for outcome evaluation [Electronic version]. *Physical Medicine And Rehabilitation: State Of The Art Reviews, 11*(2), 261-288. Available at: http://www.rasch.org/memo66.htm

APPENDIX

(A) WINSTEPS Program Code

```
; This file is CSCC Stat Teacher Eval Data
& INST
Title="Teacher Eval Run"
Name1=1
Item1=3
NI=28
;1 is strongly disagree
;2 is disagree
;3 is neutral
;4 is agree
;5 is strongly agree
Codes=12345
MODELS=R
Tables=111111111111111111111111
ASCII=Y
MUCON=0
&END
Defined Course Requirements
Defined Grading System
Knowledge of Subject
```

Enthusiastic about Subject Matter
Examples to Help Understand
Emphasized Major Points
Explained Material
Encouraged Students to Ask Questions
Answered Questions Related to Subject
Helped Those Who Needed Help
Treated Students with Respect
Maintained an Atmosphere Helpful to Learning
Used Class Time Well
Was Available the Full Class Period
Test/Assignments Returned within a Week
Would Recommend This Instructor
Text Helped me Learn
Adequate Classroom Facilities
Promoted Teacher Student Discussion
Kept Students Informed of Their Progress
Challenged Students to Think
Test were Reasonable
Seemed Well Prepared
Set High Standards for Students
Regularly Completed Homework
Regularly Attended Class
Helped Me Improve My Problem Solving Skills
Course was Appropriate to My Skill Level
END NAMES

(B) Data -- Raw Scores

```
4455354455555555455555555544
5555555555555555445555555555
53555435555555**444354554545
5555555555555555455555555555
5555544555554525345454543544
4444443444444444343344445544
4444444554444444443444444444
22*4431424434455413123435534
4544544455555554445555555444
5555554454555554554455555555
5555555555555545555555555555
545444334454445*443343345533
555555555555555555535555323*
5554544444444445444344444544
5555555555555555555555555555
5555555555555555555555555555
3343444443444344443234444423
25554545555555*1555545545545
5455555555555555555555555555
```

(C) Sample of Rating Scale Student Evaluation

Instructor's Name ________________________________
Mathematics Department

Arts and Sciences
Spring 2*** Full
Math ***
Elementary Statistics

Part I: Instructor Rating

Use a soft lead #2 pencil and express anonymously your views of the way the instructor taught this course. Rate the instructor on a scale of one (1) to five (5), with five being the highest rating.

1 = Strongly Disagree, 2 = Disagree, 3 = Neutral, 4 = Agree, 5 = Strongly Agree

The instructor clearly defined course requirements
The instructor clearly defined the grading system
.
.
The instructor was available for the full scheduled class period
Test and/or assignments were returned within a week
I would recommend this instructor to someone wanting to learn
.
.
The instructor challenged students to think
Tests were reasonable in length and difficulty
.
.
.The course helped me improve my problem solving skills