

**DETECTING MEASUREMENT DISTURBANCE EFFECTS:
THE GRAPHICAL DISPLAY OF ITEM CHARACTERISTICS**

Randall E. Schumacker
University of North Texas
(940) 565-3962
rschumacker@unt.edu

Robert Mount
Executive Director
Dallas ISD
(972) 749-5982
mount@dallasisd.org

George A. Marcoulides
California State University – Fullerton
(714) 278-7033
gmarcoulides@fullerton.edu

Paper presented at the American Educational Research Association annual meeting
Montreal, Canada
April 13, 2005
Delta Centre Ville, Salon 532
10:35am – 12:05pm

DETECTING MEASUREMENT DISTURBANCE EFFECTS: THE GRAPHICAL DISPLAY OF ITEM CHARACTERISTICS

ABSTRACT

Traditional identification of misfitting items in Rasch measurement models has interpreted the Infit and Outfit z standardized statistic. A more recent approach made possible by Winsteps is to specify “group = 0” in the control file and subsequently view the item characteristic curve for each item against the true probability curve. The graphical display reveals whether an item follows the true probability curve or deviates substantially, thus indicating measurement disturbance. Probability of item response and logit ability are easily copied to the clipboard and pasted into Microsoft Excel. An example control file, output item data, and subsequent preparation of an overlay graph for acceptable items and misfit items are presented using Winsteps and Microsoft Excel. For comparison purposes the data are also subjected to an analysis using the MD mapping procedure.

DETECTING MEASUREMENT DISTURBANCE EFFECTS: THE GRAPHICAL DISPLAY OF ITEM CHARACTERISTICS

It has been shown that the Rasch model allows for the accurate measurements of individual differences on a true linear scale (Rasch, 1960/1980; Wright, 1977; Wright & Douglas, 1977). Few other mathematical models allow for the independent estimation of person ability measures and item difficulty calibrations (Anderson, 1973; Barndorff-Nielsen, 1978; Rasch, 1961; Wright & Stone, 1979). The logistic function in the Rasch model provides for both linearity of scale and generality of measure (Wright & Stone, 1979). Georg Rasch (1960/1980) called this particular characteristic “specific objectivity.” Therefore, accurate estimates of person ability and item difficulties are possible, yet measurement disturbances must be identified and taken into consideration (Smith, 1988a, 1988b, 1991). Winsteps permits the visual display of item characteristic curves against the true probability curve. Winsteps also permits the output of item data such that item characteristic curves can be compared in Excel. A graphical display of item characteristic curves against the true probability curve in Excel provides an easy way to compare items for measurement disturbance when using Winsteps software. Similarly, MD mapping procedures also provide for a graphical display of item characteristics.

Measurement disturbances are conditions that interfere with the measurement of some underlying psychological construct. Thorndike (1949) developed a list of possible disturbances to the measurement process. Smith (1985) later classified measurement disturbances into three general categories: (a) disturbances that are the results of characteristics of the person that are independent of the items, (b) disturbances that are

the interaction between the characteristics of the person and the properties of the items, and (c) disturbances that are the results of the properties of the items that are independent of the characteristics of the person. The classification of measurement disturbances is important in that the source of the measurement disturbance dictates the techniques necessary to detect its presence. Disturbances that are characteristics of the person and independent of the items include, but are not limited to (a) start-up, (b) plodding, (c) cheating, (d) illness, (e) boredom, and (f) fatigue. Measurement disturbances associated with the interaction of the person and the properties of the items are (a) guessing, (b) item content, (c) item type, and (d) item bias. With the Rasch model, only two conditions determine the outcome of the interaction between the person and any item on a test: (a) the amount of the trait possessed by the person, and (b) the amount of the trait necessary to provide a certain response to a given stimulus (Smith, 1991a). These conditions are commonly referred to as *person ability* and *item difficulty*. Any other conditions that influence outcomes are considered measurement disturbances. Glaser (1949, 1952) and Mosier (1941) felt that a person would exhibit consistently correct answers to relatively easy items, consistently incorrect responses to difficult items, and inconsistent responses to items centered on their ability level. Since inconsistent responses could be associated with measurement disturbances, Thurstone and Chave (1929) believed that some criterion should be established so that inconsistent responses could be eliminated.

Infit Mean Square and Outfit Mean Square were provided initially to identify misfitting items to the Rasch model. Research later indicated that the z standardized statistic performed better in identifying misfitting items, i.e., $z > +/- 2.0$. In latent trait analysis, item characteristic curves have also played a role in visually displaying the item

response probability against person ability. Items with probability values greater than zero will have an item characteristic curve (ICC) that is steeper than the modeled curve and items with values less than zero will have an ICC that is flatter than the modeled curve. Items with negative fit statistics tend to have steeper observed ICCs than the true probability curve, indicating an over fit to the model, while items with positive fit statistics tend to have flatter observed ICCs than the true probability curve, indicating an under fit to the model.

METHODS AND PROCEDURES

Winsteps was used to produce calibrated item and person logit measures. There were 142 persons and 12 items that were dichotomously scored on a job satisfaction instrument. Traditional fit measures were output including the Infit and Outfit z standardized statistic. The “groups = 0” option was placed in the control file to produce individual item characteristic curves in Winsteps. In addition, Winsteps output of each items data file was sent to the clipboard. The data file was then pasted into Microsoft Excel. The Chart wizard in Excel was used to produce an overlay of item plots against the true probability curve. A graphical display was used to visually identify items with measurement disturbance effects. For comparison purposes the data were also subjected to an analysis using the MD mapping procedure (see comparison section below).

Data Analysis

The following control file was used as input into Winsteps. It reflects twelve job criteria that were coded, 1 = present and 0 = absent. The Winsteps graphical display output was used to view the item characteristic curves and copy data to the clipboard.

Item data was pasted into Excel and the Chart wizard was used to overlay item characteristic curves for acceptable and misfit items. The Winsteps control file was:

```
; This file is JOB.CON
&INST
TITLE='Job Satisfaction Criteria'
NI=12
ITEM1=10
NAME1=1
FILE = JOB.TXT
PERSON=Client
CODES=01
XWIDE=1
GROUPS=0
ITEM=criteria
PFILE=criteria.PF
IFILE=criteria.IF
&END
PERFORMANCE ADEQUACY
TIME FRAME
SATISFACTION
PHYSICALLY OR MENTALLY TIRED
HAPPINESS OF OTHERS
PRODUCED DESIRED RESULTS
PLEASED WITH PERFORMANCE
COMPLETION OF ALL STEPS
GREAT DEAL OF TIME
AWARE OF RESOURCES
FAMILY WOULD NOT BE HAPPY
SUCCESSFULLY COMPLETED
END NAMES
```

Table 1 displays the item data for items 10 and 4. The item data were copied to the clipboard in Winsteps, and then the Excel Chart wizard was used to overlay the two item characteristic curves. Item characteristic curves that did not follow the true probability curve are misfitting and indicate measurement disturbance. The type of measurement disturbance, however, will have to be decided by the researcher.

Table 1. Item Characteristic Curve Data (Items 10 and 4).

<i>Item 10</i>		<i>Item 4</i>	
<u>X1</u>	<u>Y1</u>	<u>X2</u>	<u>Y2</u>
-2.52	0.4	-5.03	0.01
-1.66	0.99	-4.17	0.01
-1.08	0.727273	-3.59	0.090909
-0.6	0.666667	-3.11	0.01
-0.17	0.99	-2.68	0.01
0.25	0.875	-2.26	0.125
0.68	0.857143	-1.83	0.142857
1.14	0.727273	-1.37	0.363636
1.66	0.761905	-0.85	0.285714
2.29	0.65	-0.22	0.4
3.21	0.555556	0.7	0.62963
4.57	0.99	2.06	0.99

RESULTS AND INTERPRETATION

Figure 1 Excel Chart plots two item characteristic curves against the true probability curve. Item 10 has Infit zstd = 7.1 and Outfit zstd = 7.3. Item 10 does not graphically follow the true probability curve (diamond legend symbol). Similarly, Item 4 has Infit zstd = .8 and Outfit zstd = .5. Item 4 is slightly better in following the true probability curve.

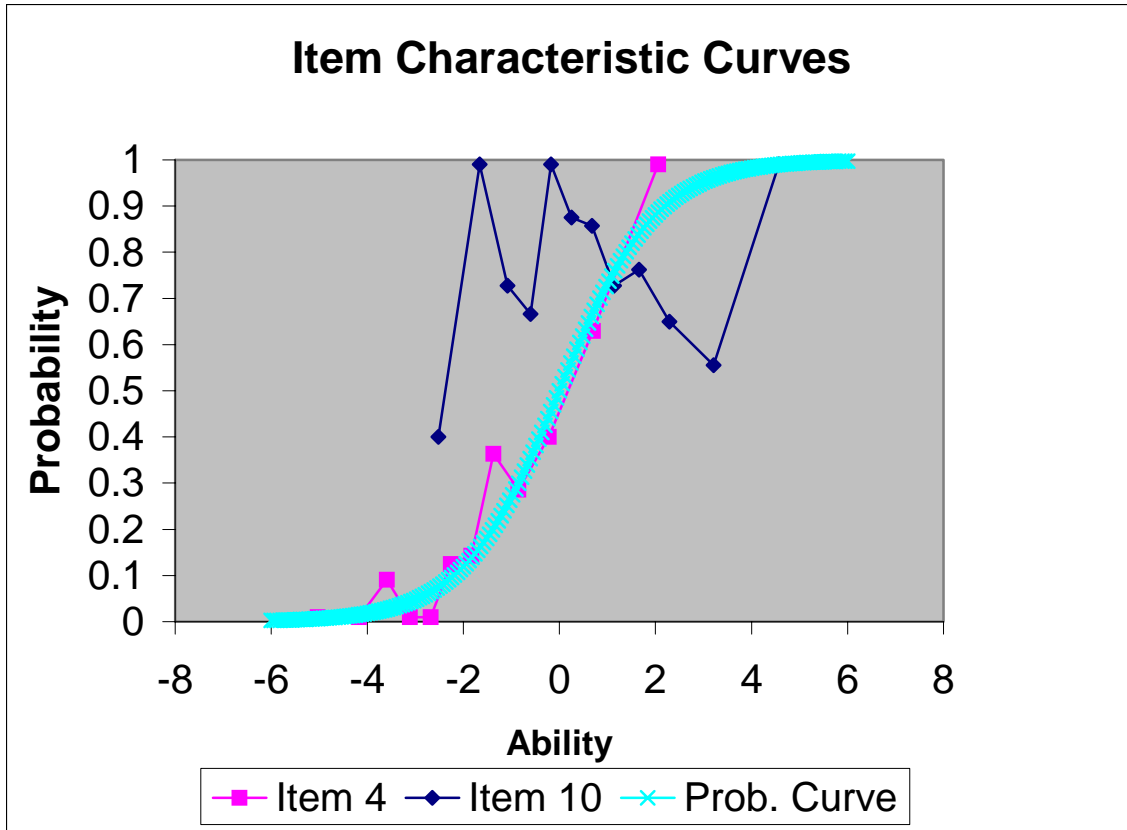


FIGURE 1. Item Characteristic Curves

Comparison of Results with the MD Model

The MD (Marcoulides & Drezner, 1993, 1995, 1997) model posits that in any measurement design with a distribution of random observations $X = (x_{ijk\dots n})$ ($i = 1, \dots, m; j = 1, \dots, p; k = 1, \dots, q; \dots, n = 1, \dots, r$) (e.g., representing m people taking a test with p items), n points (e.g., an examinee's score) are located in a dimensional space and weights (w_{ij}) between points need to be determined for $i, j = 1, \dots, n$. The weights express the importance of the proximity between points in space (e.g., the similarity in examinee ability level estimates or item difficulty). One can find

the points representing examinee ability or item difficulty by minimizing the objective function:

$$f(X) = \frac{\sum_{i,j=1}^n w_{ij} d_{ij}^2}{\sum_{i,j=1}^n d_{ij}^2}$$

where X is a vector of values for the points (defined according to the latent trait of interest – either examinee ability, item difficulty, etc.), d_{ij} is the Euclidean distance between points i and j , and weights (with $w_{ii} = 0$ for $i=1, \dots, n$ and $w_{ij} = w_{ji}$) are determined by using:

$$w_{ij} = \frac{1}{D_{ij}^p}$$

where D is the n -dimensional distance between points i and j , and the power p is a parameter that maximizes the correlation coefficient r between the vectors d_{ij} and D_{ij} for $i > j$ (for further details, see Marcoulides & Drezner, 1993; 1995; 1997). The actual values of X (i.e., the observed values on the latent trait of interest) are determined by calculating the eigenvectors of the second smallest eigenvalues of a matrix S (whose elements are defined as $s_{ij} = \sum_j w_{ij}$ and $s_{ij} = -w_{ij}$). The eigenvectors associated with the eigenvalues of S also provide coordinates of a diagnostic scatter plot (either one-dimensional or two-dimensional) for examining the various observations and conditions within a facet in any measurement design. All the eigenvalues are non-negative (for non-negative weights) with the smallest one being zero with an associated eigenvector of 1's. It is important to note that the problem is invariant under the transformation $w_{ij} = w_{ij} + c$ for any given

constant c . As such, even if there are any negative points, the problem can be solved by adding a positive constant c so that all points are non-negative.

Figure 2 presents a plot using the MD model and shows the distribution of items patterns using the example data. As can be seen in Figure 2, the patterns corroborate those observed using Winsteps. Item #10 and Item #4 appear problematic and require further study.

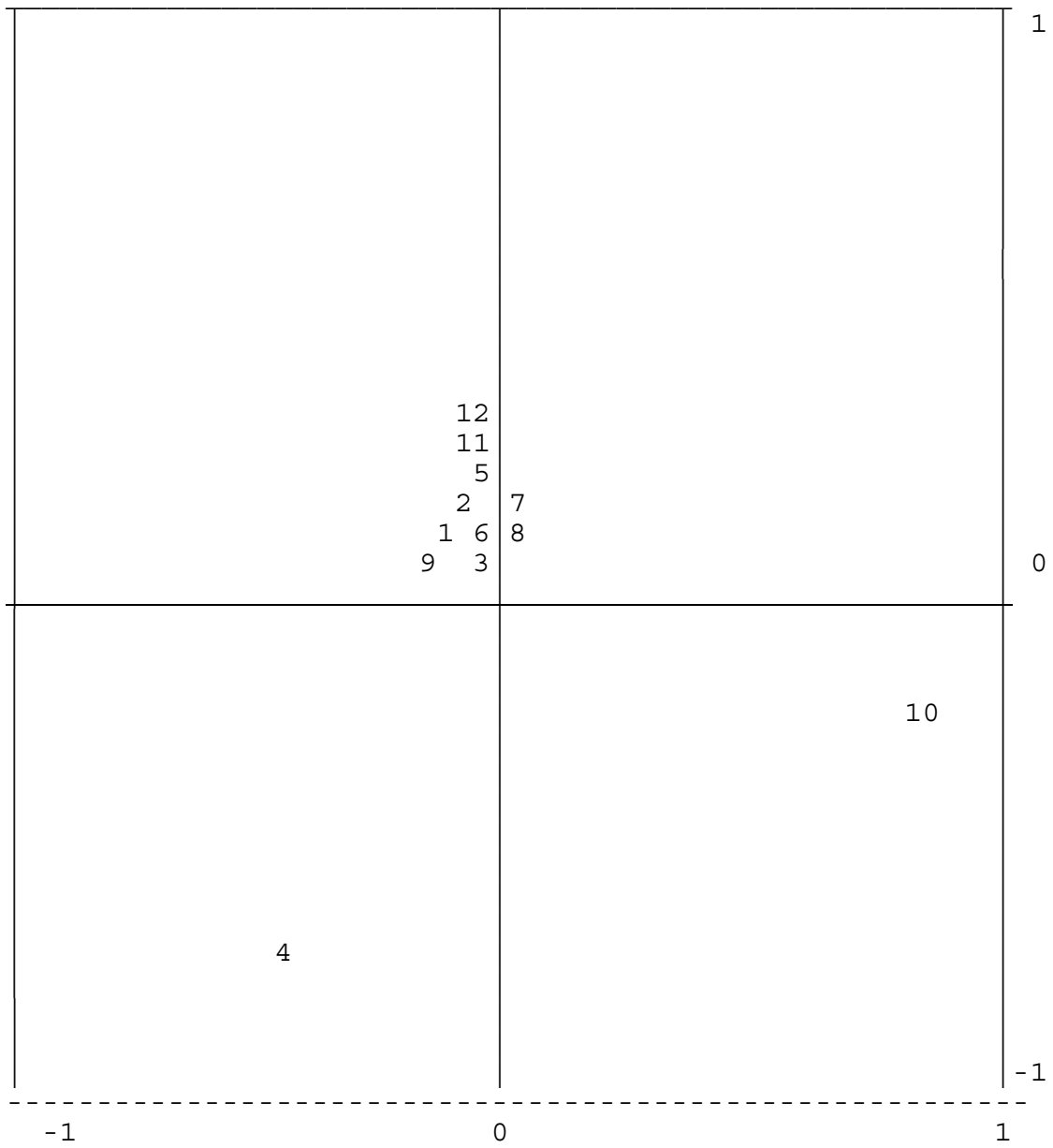


FIGURE 2. MD Model Plot of Items

CONCLUSION

Rasch measurement analysis has traditionally determined whether items were misfitting the Rasch model by examining Mean Square values (Gustafsson, 1980). Recent research has indicated that the z-standardized statistic should be used. A graphical display however provides an easy way to view items with measurement disturbance effects. Whether an item pattern reflects a true item characteristic curve is a useful way to visual find items with measurement disturbance effects. Person characteristic curves and test characteristic curves can also be generated in the same manner as outlined in this paper. In addition, the MD model also permits a graphical display of item characteristics. For comparison purposes, the MD model only contained person and item difficulty; however, other item characteristics could be added to the MD model, e.g. item formats.

REFERENCES

- Andersen, E. B. (1973). Goodness of fit test for the Rasch model. Psychometrika, 38, 123-140.
- Barndorff-Nielsen, O. (1978). Information and exponential families in statistical theory. New York: Wiley.
- Glaser, R. (1949). A methodological analysis of the inconsistency of responses to test items. Educational and Psychological Measurement, 9, 721-739.
- Glaser, R. (1952). The reliability of inconsistency. Educational and Psychological Measurement, 11, 60-64.
- Gustafsson, J-E. (1980). Testing and obtaining fit of data to the Rasch model. The British Journal of Mathematical and Statistical Psychology, 33, 205-233.
- Marcoulides, G.A., & Drezner, Z. (1993). A procedure for transforming points in multi-dimensional space to a two-dimensional representation. *Educational and Psychological Measurement*, 53(4), 933-940.
- Marcoulides, G.A., & Drezner, Z. (1995, April). A new method for analyzing performance assessments. Paper presented at the Eighth International Objective Measurement Workshop, Berkley, CA.
- Marcoulides, G.A., & Drezner, Z. (1997). A method for analyzing performance assessments. In M. Wilson, K. Draney, G. Engelhard, Jr. (Eds.). *Objective measurement: Theory into practice*. Ablex Publishing Corporation.
- Marcoulides, G.A., & Drezner, Z. (2000). A procedure for detecting pattern clustering in measurement designs. In M. Wilson, & G. Jr. Engelhard (Eds.). *Objective measurement: Theory into practice*. Stamford, CT: Ablex Publishing Corporation.
- Mosier, C.I. (1941). Psychophysics and mental test theory: II. The constant process. Psychological Review, 47, 235-249.
- Rasch, G. (1961). On general laws and the meaning of measurement in psychology. Proceedings of the Fourth Berkeley Symposium on Mathematical statistics and probability. Berkeley: University of California Press, 4, 321-333.
- Rasch, G. (1960/1980). Probabilistic models for some intelligence and attainment tests (Rev. ed.). Chicago: University of Chicago Press.

Smith, R. M. (1985). A comparison of Rasch person analysis and robust estimators. Educational and Psychological Measurement, 45, 433-444.

Smith, R. M. (1988a). A comparison of the power of Rasch total and between item fit statistics to detect measurement disturbances. A paper presented at the annual meeting of the American Educational Research Association, New Orleans.

Smith, R. M. (1988b). The distributional properties of Rasch standardized residuals. Educational and Psychological Measurement, 48, 657-667.

Smith, R. M. (1991). The distributional properties of Rasch item fit statistics. Educational and Psychological Measurement, 51, 541-565.

Thorndike, R. L. (1949). Personnel selection: Test and measurement techniques. New York: Teachers College Press.

Thurstone, L. L., & Chave, E. J. (1929). Measurement of attitudes. Chicago: University of Chicago Press.

Wright, B. D. (1977). Solving measurement problems with the Rasch model. Journal of Educational Measurement, 14, 97-116.

Wright, B. D., & Douglas, G. A. (1977). Best procedures for sample-free item analysis. Applied Psychological Measurement, 1, 281-294.

Wright, B. D., & Stone, M. (1979). Best test design. Chicago: MESA Press.