Open access • Journal Article • DOI:10.1177/014662168601000301

Graphical analysis of item response theory residuals

Larry H. Ludlow

Institutions: Boston College

Published on: 01 Sep 1986 - Applied Psychological Measurement (SAGE Publications)

Topics: Goodness of fit, Item response theory and Residual

Graphical Analysis of Item Response Theory Residuals

Larry H. Ludlow
Boston College

A graphical comparison of empirical versus simulated residual variation is presented as one way to assess the goodness of fit of an item response theory model. The two forms of residual variation were generated through the separate calibration of empirical data and data "tailored" to fit the model, given the empirical parameter estimates. A variety of techniques illustrate the utility of using tailored residuals as a specific baseline against which empirical residuals may be understood.

This paper presents an analytic method for isolating and identifying departures from the fit of an item response theory (IRT) model. The specific techniques employed focus on the graphical comparison of empirical residual variation to baseline residual variation. The baseline variation is the result of data generated to fit the model, given the empirical parameter estimates. The baseline residuals thus serve as the reference background for interpreting the empirical residuals. Although the Rasch model is applied in this paper, the principles that are discussed and illustrated hold for the residual analysis of any IRT model.

The Model

The Rasch model has been developed for a wide range of testing conditions (e.g., Andrich, 1978; Embretson, 1984; Fischer, 1973; Masters, 1982; Rasch, 1960; Whitely, 1980). Masters and Wright (1984) brought together the fundamental structure incorporated by five of the more frequently applied models. The rating scale model (Andrich, 1978) employed in the present application takes the form

$$\pi_{nix} = \frac{\exp\left\{\sum_{j=0}^{X_{ni}}[\beta_n - (\delta_i + \tau_j)]\right\}}{\sum_{k=0}^{m}\exp\left\{\sum_{j=0}^{k}[\beta_n - (\delta_i + \tau_j)]\right\}}, \qquad (1)$$

where π_{nix} is the probability of observing the score x ($x = 0,1, \ldots, m$),
β_n is the performance parameter for person n ($n = 1,2, \ldots, N$),
δ_i is the difficulty parameter for item i ($i = 1,2, \ldots, L$), and
τ_j is the difficulty parameter for the $j = 1,2, \ldots, m$ category thresholds.

Several Rasch model parameter estimation techniques have been proposed (e.g., Wright & Masters, 1982). Regardless of the technique, the minimally sufficient statistics for the rating scale model are simply the person, item, and category total scores. Once β_n, δ_i, and τ_j are estimated, they are used to compute the expected response for every person on each item. These expected responses are compared to the observed responses; their difference is a residual.

APPLIED PSYCHOLOGICAL MEASUREMENT
Vol. 10, No. 3, September 1986, pp. 217–229

0146-6216/86/030217-13$1.90

Under the model, the expected response for a person taking any item is

$$E_{ni} = \sum_{k=0}^{m} k\pi_{nik} \quad . \tag{2}$$

The expected responses have variance

$$W_{ni} = \sum_{k=0}^{m} (k - E_{ni})^2 \pi_{nik} \quad . \tag{3}$$

Estimated residuals may be expressed in standard form as

$$Z_{ni} = \frac{X_{ni} - E_{ni}}{(W_{ni})^{1/2}} \quad . \tag{4}$$

These residuals have an expected value of 0 and a variance of 1.

Graphical Residual Analysis

The statistical literature provides extensive discussion of the graphical representation and analysis of residual variation. For example, Anscombe and Tukey (1963), Barnett and Lewis (1978), Draper and Smith (1981), and Cook and Weisberg (1982) offered many practical reference patterns, analytic strategies, and techniques for plotting and interpreting residuals. An analogous methodology has not firmly established itself in the IRT literature. There is not, for example, any discussion of graphical methods of residual analysis in Lord and Novick (1968), Lord (1980), Hulin, Drasgow, and Parsons (1983), Wainer and Messick (1983), or Weiss (1983).

Techniques for the graphical analysis of IRT residuals have demonstrated their diagnostic utility (Hambleton & Murray, 1983; Mead, 1975; Wright & Masters, 1982; Wright & Stone, 1979). An appreciation of the utility of graphical techniques in general, however, has been lacking. This is due, in part, to the lack of a technology and rationale for obtaining baseline, boundary-defining patterns of expected residual variation. Baseline patterns representing expected residual variation are essential in a graphical analysis, because expected or unexpected patterns formed by a scatter of points are not always evident. Hahn and Shapiro (1967, Chap. 8) and Daniel and Wood (1980, Chap. 3), for example, demonstrated how probability and cumulative distribution plots of randomly generated normal deviates can vary drastically from one simulation to the next. Furthermore, not only are graphical analyses subject to different interpretations by different individuals, but an individual's response may differ from occasion to occasion (Collet & Lewis, 1976). Graphical residual analyses, consequently, require the generation of baseline configurations against which the plot of interest may be compared and interpreted.

At present, graphical IRT residual analysis research concentrates on two approaches to generating data and baseline patterns. In the first, parameters are sampled from hypothetical distributions. Data are then generated to fit the model, given these parameters. This method is useful for exploring the effect of test and model characteristics upon the distribution of residuals. A series of simulation studies by Ludlow (1983) illustrated some structural patterns to expect from Rasch model residuals under a variety of hypothesized testing conditions.

The second approach begins with the calibration of empirical data. If the empirical estimates are accurate, then the residuals should form predictable patterns. To determine how those patterns should appear, however, data known or "tailored" to fit the model must be generated. This is accomplished by employing the empirical estimates as data generating parameters. The calibration of these tailored data yields residual variation known to fit the model.

Residuals produced by the tailored method provide the relevant framework for revealing deviations from the model in the empirical data. If the empirical data fit the model, the empirical and tailored data should yield residuals which behave similarly. The hypothetical simulations establish the broad background for what is possible. The tailored simulations focus on the empirical data and define its particular baseline.

Since each tailored simulation is one "what if" event, it is necessary to replicate simulations in order to minimize capitalizing on the chance generation of atypical data. The replication issue is problematic, as it requires determination of both the number of tailored analyses to perform, and how the multiple analyses are to be compared to one another. Experience suggests that three to five

sets of tailored analyses may prove sufficient for uncovering major irregularities.

Application

The data are from the administration of an instrument constructed to measure attitudes toward blindness (Courington, Lambert, Ludlow, Wright, & Becker, 1983). There were $L = 19$ items, $N = 222$ persons, and $M = 4$ response categories. The items were scored: Strongly Agree = 3, Agree = 2, Disagree = 1, Strongly Disagree = 0 (i.e., the higher the score, the more positive the attitude). Three interviewers collected the data from blind patients scheduled to participate in a rehabilitation program at the United States Veterans Administration Hines Hospital, Hines, Illinois.

Table 1 contains the empirical results, listed in their calibration order. "Value" is the item calibration, "SE" is the standard error of estimate, and "Fit" is a goodness of fit statistic computed from the vector of standardized residuals (Wright & Masters, 1982). It may be represented as the normal (cube root) transformation of a weighted mean square,

$$t_i = (V_i^{1/3} - 1)(3/q_i) + (q_i/3) \quad , \tag{5}$$

where the weighted mean square is

$$V_i = \sum_n^N W_{ni} Z_{ni}^2 \Big/ \sum_n^N W_{ni} \quad . \tag{6}$$

The expected value of V_i is 1 and its expected variance is

$$q_i^2 = \sum_n^N (C_{ni} - W_{ni}^2) \Big/ \left(\sum_n^N W_{ni} \right)^2 \quad , \tag{7}$$

where C_{ni} is the kurtosis of X_{ni}.

In the present application, a positive residual indicates a response more favorable than expected under the model. A negative residual indicates a response more unfavorable than expected. A positive fit statistic results from any combination of such inconsistent, unexpected responses. A negative fit statistic, however, results from residual variation that is less than that expected under the model. For these data, a negative fit means that an item did not provoke many Strongly Agree or Strongly Disagree responses.

Given the results in Table 1, only a respondent with a very positive attitude would be expected to strongly agree with the statement "Being blind is an asset to marriage". Only a respondent with a very negative attitude would be expected to strongly

Table 1
Rating Scale Model Calibration Results

Item	Description	Value	SE	FIT
19	asset to marriage: "caring"	2.35	.11	0.03
13	better telephone work: "sensitive"	1.08	.11	-2.33
15	more "honest" than sighted	0.92	.11	0.99
16	can endure boring tasks: "patient"	.56	.11	-0.75
4	don't superficially "judge"	.55	.11	0.01
17	closer to spouse: "closeness"	.53	.11	0.09
10	complain less: "accepting"	.37	.11	-1.49
14	distracted less: "focused"	.32	.11	.63
12	"understand" feelings better	.18	.11	-0.80
11	"loyal" friend	-0.34	.12	-0.83
18	good "supervisor"	-0.35	.12	-1.73
9	good "negotiator"	-0.60	.12	-4.69
5	"participate" in activities	-0.62	.12	0.21
7	develop extra "sense"	-0.66	.12	3.36
6	superior piano "tuner"	-0.68	.12	-0.24
8	good social worker: "therapist"	-0.70	.12	-3.22
2	offer "spouse" satisfactory sex	-0.75	.12	3.45
1	"work" as well as anyone	-1.08	.12	2.69
3	raise a normal child: "parent"	-1.09	.12	1.87

disagree with the statement "A blind person can raise a normal child". The threshold estimates ($\hat{\tau}_1 = -2.53$, $\hat{\tau}_2 = -0.17$, $\hat{\tau}_3 = 2.70$) indicate that most responses were recorded as Agree or Disagree. Overall, the item and threshold orders are sensible and conform to the intent of the instrument. The positive fit statistics, however, indicate the presence of numerous responses inconsistent with the model.

The following residual analysis used three tailored simulations. Each simulation, starting from the same empirical estimates, generated data to fit the model for 222 patients responding to 19 questions with four choice options. Each of the simulations was calibrated separately, residuals analyzed, results compared to the empirical results, and then compared to one another. The results from one simulation, as opposed to all three intermingled, are presented.

Probability Distributions

This analysis begins with the observation that a $N \times L$ matrix of residuals may be analyzed from virtually an infinite variety of perspectives. Nevertheless, one reasonable first step is to simply assess the distributional properties of the residuals. Under the model, the standardized residuals are assumed $N(0,1)$. Thus, a comparison of empirical and tailored residual normal probability plots may highlight gross distributional differences.

Separate normal probability plots (Blom, 1958) were constructed and compared. They appeared to show a greater than expected concentration of large negative empirical residuals. In order to form a clearer impression of the differences, a quantile-quantile (Q–Q) plot of the ordered empirical and tailored residuals was constructed (Chambers, Cleveland, Kleiner, & Tukey, 1983). If the distributions had been identical, then an identity line would have resulted. That was not the case. There was a heavier concentration of large negative empirical residuals. Finally, since the Q–Q representation of the differences was rather coarse for 4,218 residuals, a Tukey sum-difference graph was constructed (Cleveland, 1985).

In Figure 1, the abscissa represents the sum of each ordered pair of empirical and tailored residuals. The ordinate represents the difference between the residuals in each pair. Each point represents 1/100th of the total distribution, hence the smooth graph. If the two sets of residuals had the same distribution, the resulting pattern of variation would fluctuate slightly around the ordinate position of zero. However, a striking pattern emerges.

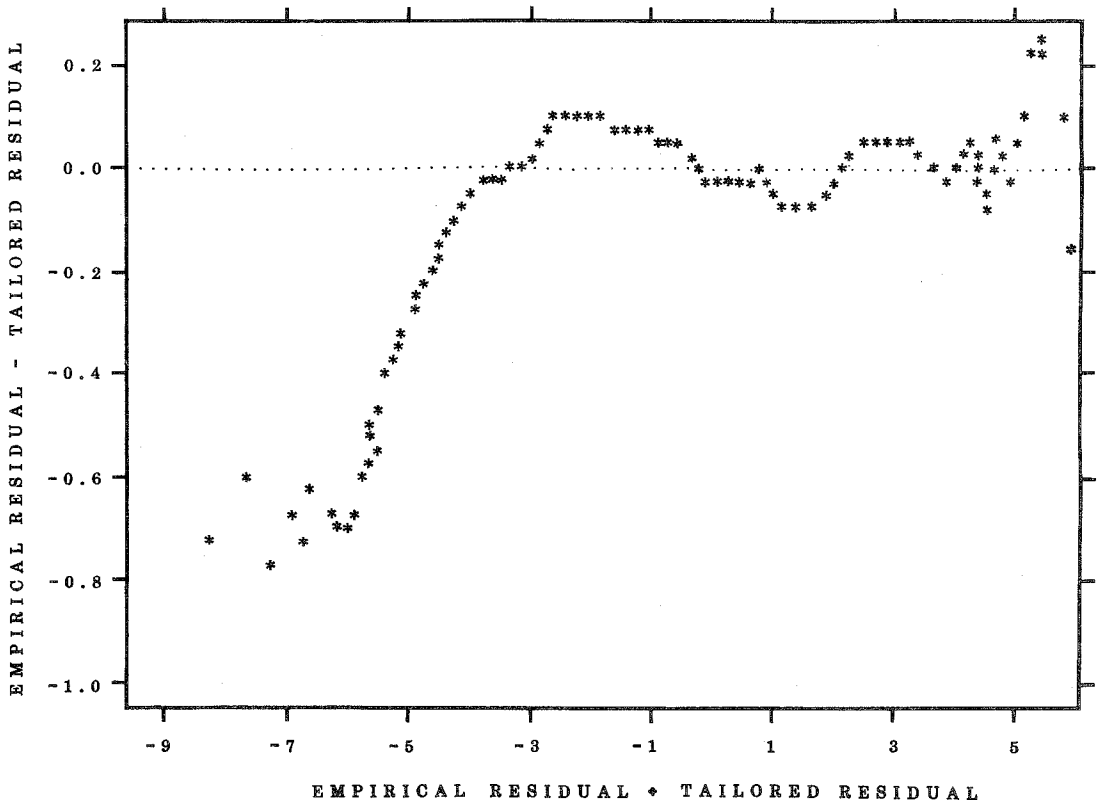
Beginning in the lower left section, there are too many large negative empirical residuals. The pattern rises and peaks, reflecting too few small negative residuals. It then drops, reflecting too few small positive residuals. The pattern then flattens somewhat, except at the extreme right where there is a tendency for some of the largest positive empirical residuals to be greater than their tailored counterparts. These distributional analyses suggest that particular attention should be focused on the large negative residuals.

Density Differences

Figure 2 reveals the density of the empirical residuals when plotted against the item calibrations. The figure represents a theme called a "graphic rational pattern" or GRP (Bachi, 1968). Each box contains GRP squares that represent the precise number of residuals found within the enclosed area. Although the practical construction of this representation is relatively complex, it offers a simple interpretation: The darker the box, the denser the pattern.

As expected under the model, the density is greatest near the ordinate position of zero. Any two columns, however, may not be directly comparable, because some columns represent the residuals from a single item while others represent the residuals from two items with similar difficulty estimates. Nevertheless, the outline of a structural skewness can be seen (from the lower left to the upper right). This structural skew necessarily results when item estimates become increasingly extreme and some people continue to respond unexpectedly. The range in estimates is relatively narrow, else the skew could have been more pronounced. The point here is that this pattern is ex-

Figure 1
A Tukey Sum-Difference Plot of the Ordered Pairs of Empirical and Tailored Standardized Residuals



pected to occur. The problem is to define a relevant baseline against which this pattern may be interpreted.

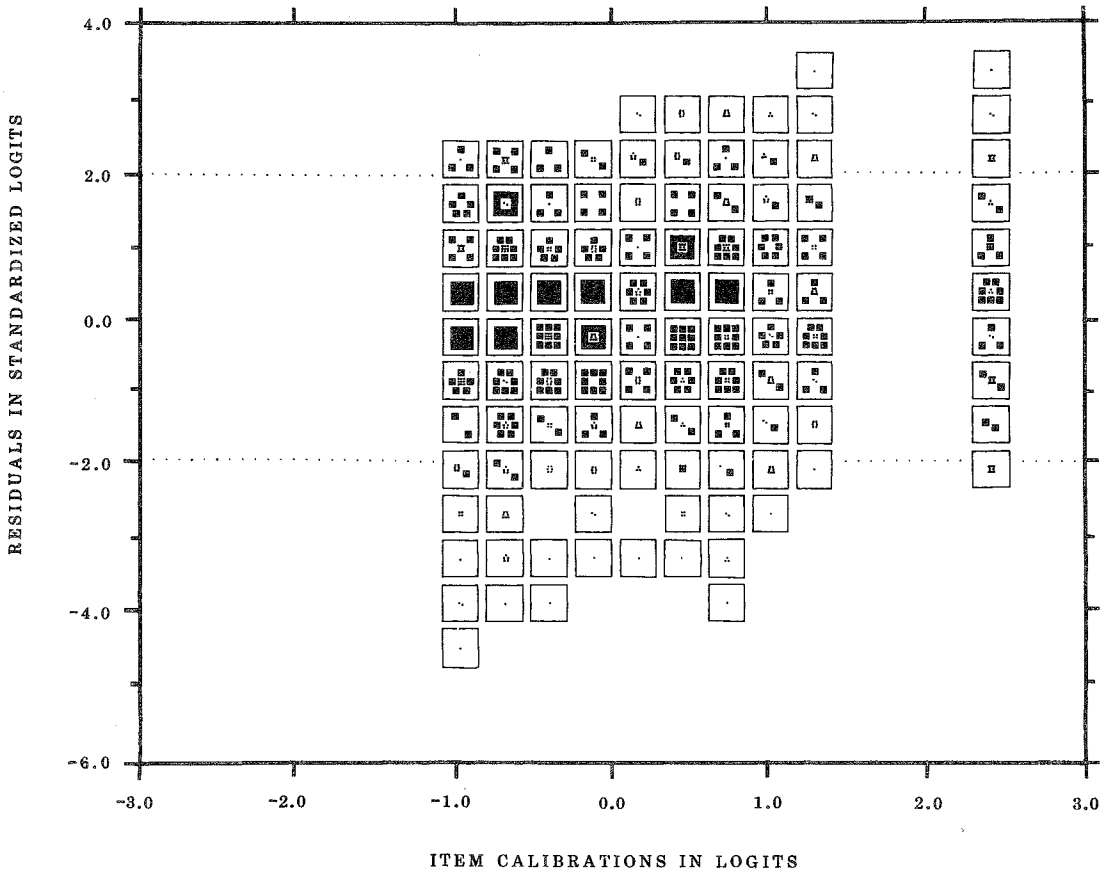
A companion plot was constructed for the tailored residuals. Instead of visually comparing the two plots, it is simpler to create a density difference plot (Chambers et al., 1983). Figure 3 shows where the empirical density is greater than the tailored density. The area of interest lies in the lower left section. This area indicates that an unexpected number of surprising Strongly Disagree responses occurred for items which had seemed likely to provoke Agree responses.

The empirical and tailored residuals were next plotted against the patient attitude estimates. The density difference plot (not shown) revealed that the large negative residuals in Figures 2 and 3 were from midrange-attitude patients. Like the previous figures, a structural skew was observed, but it was in the opposite direction.

Test Characteristic Differences

Figure 4 presents, in item sequence order, modified box-plots (the box enclosing the interquartile range has been eliminated and the median is represented by a dot; Tufte, 1983) of the tailored residuals and those empirical residuals outside the tailored range (“*”). The lower left half of the figure reveals that an unexpected number of surprising Strongly Disagree responses occurred within the first section of the instrument.

Figure 2
Density of the Empirical Standardized Residuals
Distributed Across the Item Difficulty Estimates
(Each GRP Unit Within the Boxes Represents Either 1, 10, or 100 Residuals)

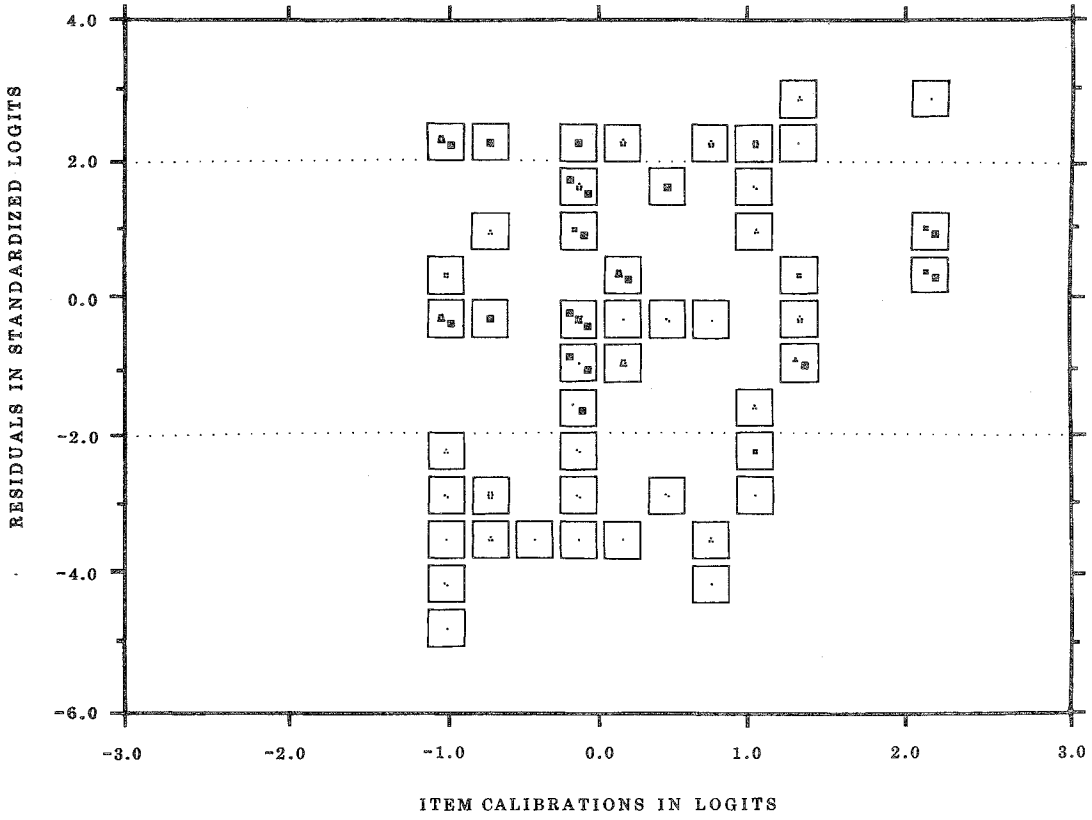


Up to this point, the analyses have isolated a group of midrange-attitude patients who provided unexpected Strongly Disagree responses on some of the first and easiest items. This situation suggested that a "start-up" effect influenced the measurement process. It is a common practice to present participants with a few trial items. None were included on this instrument. This observation raised a question about whether the surprising responses were due to the patients' unfamiliarity with the response instructions or to inconsistent judgments exercised by interviewers in recording responses.

Discussions with interviewers revealed that many patients did not initially use the suggested labels. Interviewers admitted that subjective decisions for coding those responses occurred and rested on secondary cues and hypothesized patient response styles. The interviewers claimed that they were eventually able to interpret each patient's response style, but each interviewer handled the interpretive problem in an idiosyncratic fashion. In fact, the interviewers generally adopted one of two recording strategies: conservative (primarily Disagree or Agree responses), or liberal (frequent Strongly Disagree or Strongly Agree responses).

An example of the interviewer effect is illustrated in Figure 5. The three pairs of box-plots (Tukey, 1977) represent the tailored and empirical

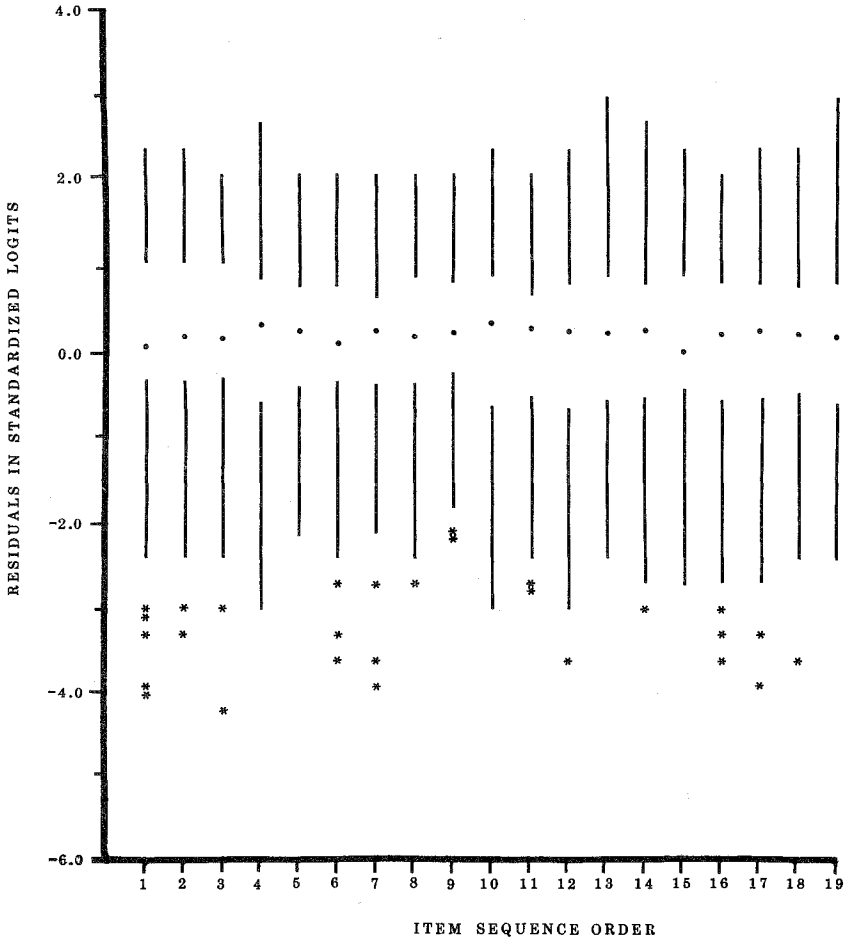
Figure 3
Density Difference Between the Empirical and Tailored Standardized Residuals
(Each Box Indicates the Area and Count of Where
the Empirical Density Was Greater Than the Tailored Density)



residual distributions on the first item (''work'') broken down by each interviewer. Consistent with their self-assessment of their liberal response-recording strategy, the residuals attributable to Interviewers A and B reflect an unexpected number of Strongly Disagree responses. Interviewer C's pattern is slightly less variable than expected, a reflection of his conservative scoring strategy. These patterns were typical for items in the first section of the instrument. When the residuals attributable to each interviewer were plotted across time, Interviewers A and B showed a tendency toward decreased residual variation, while Interviewer C showed a tendency toward increased variation. This start-up effect was a source of measurement error that led to the introduction of practice items and standardization of interviewing technique.

Figure 6 is a dot chart (Cleveland, 1984) that reveals an overall differential interviewer effect. Each dot's location was computed by first taking the difference between the mean empirical and mean tailored standardized residual for each interviewer on each item. The difference between these mean differences was then computed for each pair of interviewers. This particular process allows a comparison of interviewer-recorded responses, given the responses their patients were expected to have provided under the model. The negative vectors

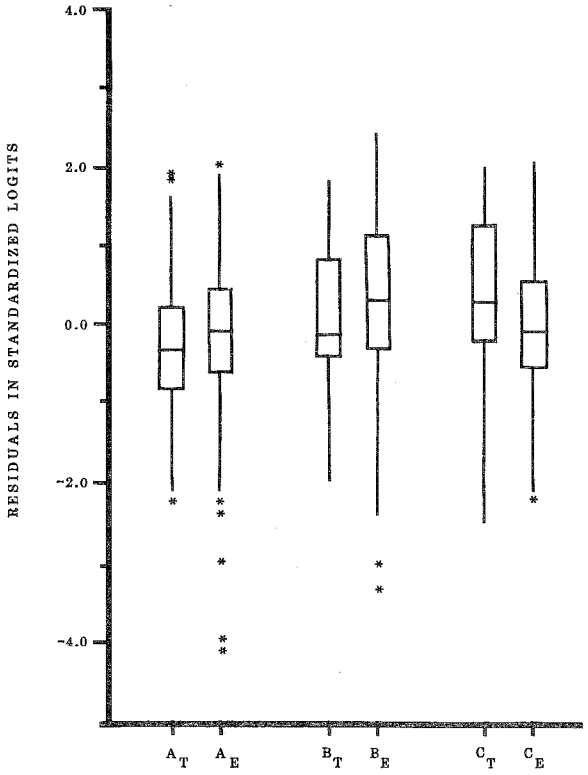
Figure 4
Box-plots of the Empirical and Tailored Standardized Residuals in Their Item Sequence Order
(Each · Represents the Median Tailored Residual; the Gap Around Each Median Represents the Interquartile Range; the Vertical Lines Extend to the Adjacent Values; Each * Represents an Empirical Residual That Fell Beyond the Tailored Distribution)



indicate items for which Interviewer B's mean recorded response was lower than that recorded by Interviewer A. The positive vectors indicate items for which Interviewer B's mean recorded response was higher than that recorded by Interviewer A.

All the mean differences should lie close to the vertical line of origin. As can be seen, there are a number of relatively large discrepancies between these two interviewers. At the bottom of the chart, the large positively oriented item (''work'') reflects the start-up effect previously discussed. The significant pattern of interest lies in the upper left section of the figure. The three largest negatively oriented items are the only ones that address some

Figure 5
Box-plots of the Empirical and Tailored Standardized Residuals on One Item (''Work'') According to the Interviewer (A, B, or C) Who Collected the Data
(The Tailored Distributions Are Noted by T, the Empirical by E; the Plots Reveal the Respective Pairs of Medians, Interquartile Ranges, and Frequency of Outside Values)



aspect of spousal relationship: ''caring'' (blindness is an asset to marriage), ''spouse'' (a blind person can offer their spouse satisfactory sex), and ''closeness'' (a blind person is closer to his spouse than a sighted person). These related items provoked an unexpected number of Strongly Disagree responses when Interviewer B, a woman, collected the data. A similar configuration resulted when her data were plotted against the other (male) interviewer.

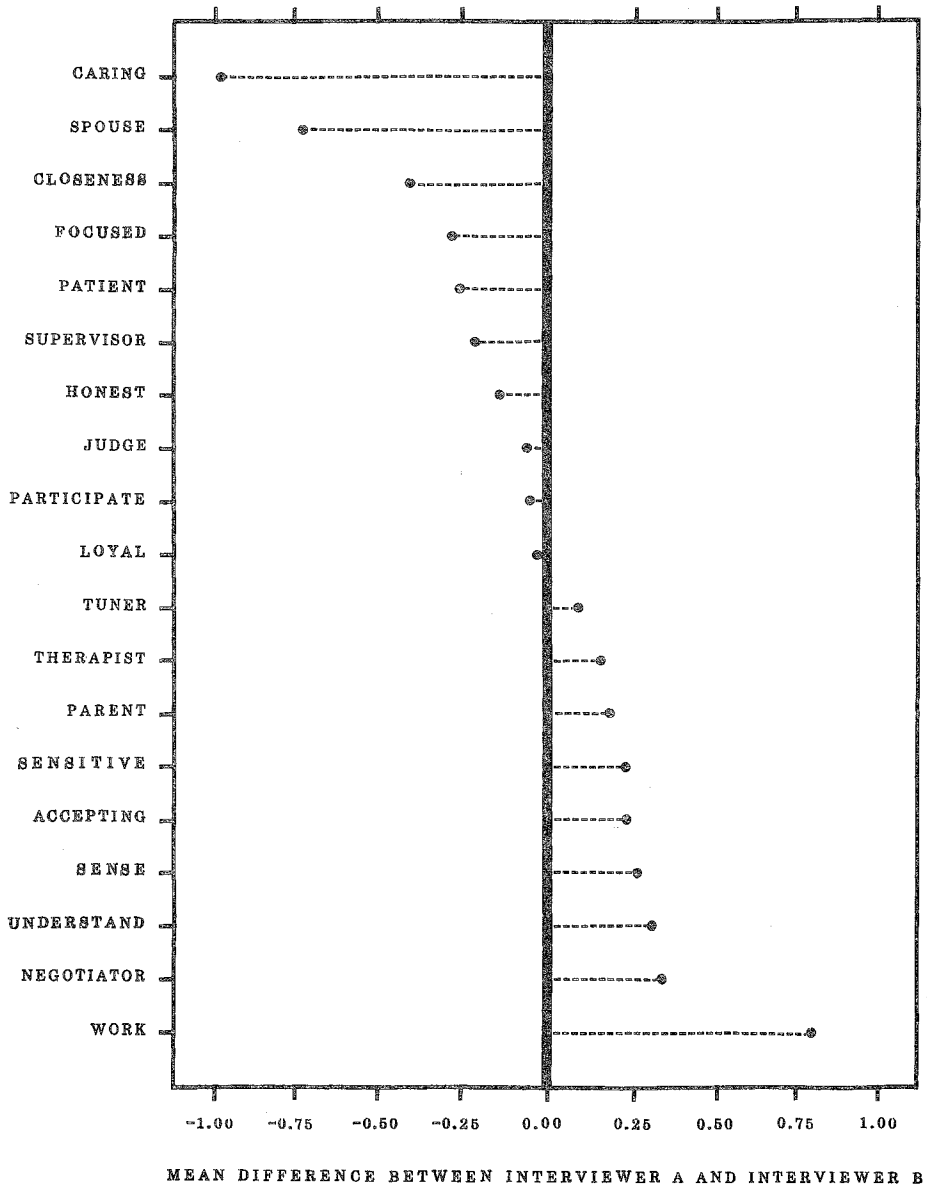
Further investigation revealed that 39% of the men interviewed by the woman had already entered the hospital (the interviews were to have been done by telephone prior to admittance). These interviews were conducted by her in each patient's private room. Of the 11 largest negative residuals for these items across all interviewers, 7 came from her ''bedroom'' interviews. The patterns in this analysis, as well as anecdotal evidence from rehabilitation staff, led to a change in interviewing locale.

Structural Differences

The problems in these data suggested that a more global assessment of misfit might be informative. The concern was with the unidimensionality of the instrument. Accordingly, the principal components of the inter-item residual correlation matrix were extracted. The two-component solution for the tailored residuals produced, as expected, a random scatter of items centrally located about the origin.

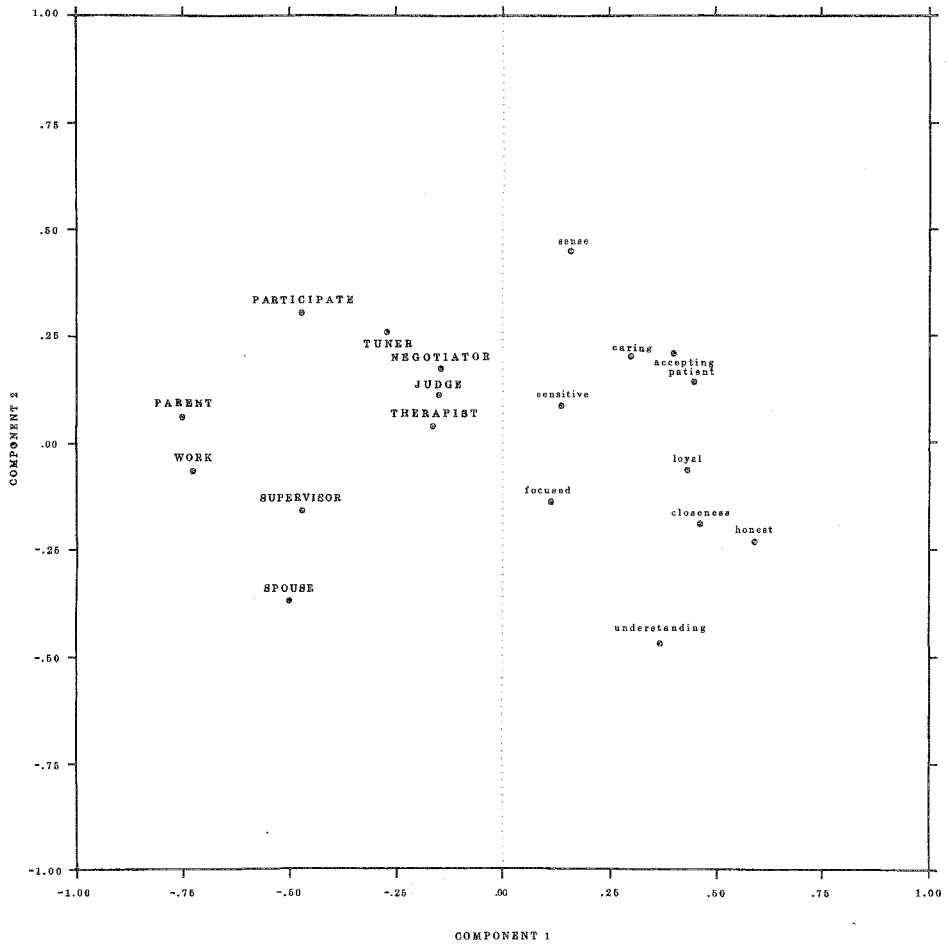
Figure 7 contains the unrotated principal com-

Figure 6
A Dot Chart of the Differential Interviewer Effect
(The Dots Represent the Result of Comparing Interviewer B's Mean Residuals to Interviewer A's Mean Residuals; Negative Vectors Translate Into Unexpected Low Responses to B; Positive Vectors Represent Unexpected High Responses to B)



ponents solution for the empirical residuals. The item loadings are identified by the keywords identified in Table 1. Both a Q–Q plot of the empirical and tailored eigenvalues (Gnanadesikan, 1977) and the simple difference between the first pair of eigenvalues (empirical $\lambda_e = 2.96$, tailored $\lambda_t = 1.63$)

Figure 7
Unrotated Two-Component Solution
for the Empirical Standardized Residuals Inter-item Correlation Matrix



suggest that a linear structure remains in the correlation matrix. Along the first component, the marginal distribution suggests that the items form two clusters. In the negative direction, the items generally address skills or roles that a blind person might be able to perform as well as or better than a sighted person (''role'' items). In the positive direction, the items generally address personal characteristics blind persons might acquire as a consequence of blindness (''personal'' items).

The presence of two subscales was supported in the final analysis. The items were separated into two groups depending on whether their first loading was negative (role) or positive (personal). Separate calibrations of the empirical and tailored data were then obtained. Finally, for each calibration the pairs of role and personal patient measures were plotted.

For the tailored data, 5 patients (2% of the sample) fell beyond the 95% confidence band. For the empirical data, a total of 31 patients (14%) fell outside the confidence band. One group of patients agreed with the personal items but disagreed with the role ones. The other group of patients agreed with the role items but disagreed with the personal

ones. For those two groups of patients, this instrument did not define a unidimensional continuum. An overall estimate of attitude was not appropriate; separate estimates were required.

Conclusions

A comparative analysis of the empirical and tailored residuals from an IRT model calibration can serve as one method for assessing the fit of a model or for comparing the fit of alternative models. In the present application, this approach revealed numerous sources of measurement error that were addressed by improving interviewing techniques, response scoring, item phrasing, patient attitude score reporting, and instrument administration. These practical problems were detected because residuals from tailored simulations provided a specific frame of reference against which deviations in the empirical variation could be discerned.

Although there is a degree of subjectivity involved whenever a graphical representation is analyzed, a knowledge of the expected forms of baseline patterns is one means of reducing the tendency to overinterpret the data. Furthermore, a hierarchical search strategy helps focus the analysis on important characteristics of the data (Ludlow, 1985). Nevertheless, as with any statistical tool, extensive experience may be required before the potential gains and pitfalls of a graphical approach are fully appreciated. This is particularly true when residual distributions from different IRT models are compared, because measurement error uncovered by one model may manifest itself in a different form in the fitting of an alternative model.

References

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

Anscombe, F. J., & Tukey, J. W. (1963). The examination and analysis of residuals. *Technometrics, 5,* 141–160.

Bachi, R. (1968). *Graphical rational patterns*. New York: Israel Universities Press.

Barnett, V., & Lewis, T. (1978). *Outliers in statistical data*. New York: Wiley.

Blom, G. (1958). *Statistical estimates and transformed beta variables*. New York: Wiley.

Chambers, J. M., Cleveland, W. S., Kleiner, B., & Tukey, P. A. (1983). *Graphical methods of data analysis*. Boston: Duxbury Press.

Cleveland, W. S. (1984). Graphical methods for data presentation: Full scale breaks, dot charts, and multibased logging. *The American Statistician, 38,* 270–280.

Cleveland, W. S. (1985). *The elements of graphing data*. Monterey CA: Wadsworth.

Collet, D., & Lewis, T. (1976). The subjective nature of outlier rejection procedures. *Applied Statistician, 25,* 228–237.

Cook, R. D., & Weisberg, S. (1982). *Residuals and influence in regression*. New York: Chapman and Hall.

Courington, S. M., Lambert, R. W., Ludlow, L. H., Wright, B. D., & Becker, S. W. (1983). The measurement of attitudes toward blindness and its importance for rehabilitation. *International Journal of Rehabilitation Research, 6,* 67–72.

Daniel, C., & Wood, F. S. (1980). *Fitting equations to data* (2nd ed.). New York: Wiley.

Draper, N. R., & Smith, H. (1981). *Applied regression analysis* (2nd ed.). New York: Wiley.

Embretson, S. (1984). A general latent trait model for response processes. *Psychometrika, 49,* 175–186.

Fischer, G. H. (1973). Linear logistic test model as an instrument in educational research. *Acta Psychologica, 37,* 359–374.

Gnanadesikan, R. (1977). *Methods for statistical data analysis of multivariate observations*. New York: Wiley.

Hahn, G. J., & Shapiro, S. S. (1967). *Statistical models in engineering*. New York: Wiley.

Hambleton, R. K., & Murray, L. (1983). Some goodness of fit investigations for item response models. In R. K. Hambleton (Ed.), *Applications of item response theory* (pp. 71–94). Vancouver, British Columbia: Education Research Institute of British Columbia.

Hulin, C. L., Drasgow, F., & Parsons, C. K. (1983). *Item response theory*. Homewood IL: Dow Jones-Irwin.

Lord, F. M. (1980). *Applications of item reponse theory to practical testing problems*. Hillsdale NJ: Lawrence Erlbaum.

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Reading MA: Addison-Wesley.

Ludlow, L. H. (1983). *The analysis of Rasch model residuals*. Unpublished doctoral dissertation, University of Chicago.

Ludlow, L. H. (1985). A strategy for the graphical representation of Rasch model residuals. *Educational and Psychological Measurement, 45,* 851–859.

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika, 47,* 149–174.

Masters, G. N., & Wright, B. D. (1984). The essential

process in a family of measurement models. *Psychometrika, 49,* 529–544.
Mead, R. J. (1975). *Analysis of fit to the Rasch model.* Unpublished doctoral dissertation, University of Chicago.
Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests.* Copenhagen: Denmarks Paedagogiske Institute.
Tufte, E. R. (1983). *The visual display of quantitative information.* Cheshire CT: Graphics Press.
Tukey, J. W. (1977). *Exploratory data analysis.* Reading MA: Addison–Wesley.
Wainer, H., & Messick, S. (Eds.). (1983). *Principals of modern psychological measurement.* Hillsdale NJ: Lawrence Erlbaum.
Weiss, D. J. (Ed.). (1983). *New horizons in testing.* New York: Academic Press.
Whitely, S. E. (1980). Multicomponent latent trait models for ability tests. *Psychometrika, 45,* 479–494.
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.

Acknowledgments

Thanks are due to Benjamin D. Wright, the editor, and Susan Catalini. Special appreciation is expressed for the advice of an anonymous reviewer. This research was sponsored in part by the USVA Hines Hospital, Rehabilitation Research and Development Center, Hines, Illinois, U.S.A.

Author's Address

Send requests for reprints or further information to Larry H. Ludlow, Boston College, School of Education, Campion Hall, Chestnut Hill MA 02617, U.S.A.