DOCUMENT RESUME

ED 390 941 TM 024 581

AUTHOR Linacre, John Michael
TITLE The Effect of Misfit on Measurement.
PUB DATE Apr 95
NOTE 8p.; Paper presented at the Annual Meeting of the International Objective Measurement Workshop (8th, Berkeley, CA, April 1995).
PUB TYPE Reports - Evaluative/Feasibility (142) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Children; *Goodness of Fit; *Item Response Theory; *Measurement Techniques; *Reliability; Simulation; *Test Items
IDENTIFIERS Calibration; Knox Cube Test; Person Fit Measures; *Rasch Model

ABSTRACT

The effects on Rasch measurement of both response underfit (noise) and overfit (mutedness or superuniformity) are described and illustrated. Misfit is identified by mean-square fit statistics. Person separation and reliability are shown to be deceptive indicators of measurement effectiveness when some items exhibit marked overfit. Theoretical considerations are confirmed by adding simulated items of known fit to a core of empirically observed Knox Cube Test responses of 34 children. Measure effectiveness, model fit, and utility must be considered together when deciding which subset of data is most informative for the particular task at hand. One subset of data may be most informative for item calibration, and another for reporting person measures based on the precalibrated items. (Contains four figures and two references.) (Author/SLD)

The Effect of Misfit on Measurement

By

John Michael Linacre
University of Chicago

Paper presented at the
Eighth International Objective Measurement Workshop

Berkeley, California
April 1995

MESA Psychometric Laboratory
Department of Education
University of Chicago
5835 S. Kimbark Ave
Chicago IL 60637-1609

Tel: (312) 702-1596, FAX (312) 702-0248
E-mail: MESA@uchicago.edu

The Effect of Misfit on Measurement

Abstract:
The effect on Rasch measurement of both response underfit (noise) and overfit (mutedness, superuniformity) are described and illustrated. Misfit is identified by mean-square fit statistics. Person separation and reliability are shown to be deceptive indicators of measurement effectiveness when some items exhibit marked overfit. Theoretical considerations are confirmed by adding simulated items of known fit to a core of empirically observed Knox Cube Test responses.

Text:
Measurement using the Rasch model requires some degree of stochasticity in the observations. For dichotomous data, one form of the model is:

$$\log\left(\frac{P_{ni}}{1 - P_{ni}}\right) \equiv B_n - D_i$$

where P_{ni} is the probability of a "correct" answer by person n on item i. B_n is the ability of person n. D_i is the difficulty of item i.

When the data fit the model, they can be expressed as

$$\{X_{ni}\} = \{P_{ni} \pm \sqrt{P_{ni}(1 - P_{ni})}\}$$

where X_{ni} is the response by person n to item i.

A t-test of the unexpectedness of any particular observation is

$$t = \frac{X_{ni} - P_{ni}}{\sqrt{P_{ni}(1 - P_{ni})}}$$

Mean-square Fit Statistics

Measurement under Rasch model conditions requires stochasticity, randomness in the data, and also that the randomness be distributed evenly across the data. Under these conditions, the data fit the model. In practice, data never fit the model perfectly, so there is either a local excess of stochasticity, "noise", or a local deficiency of stochasticity, "mutedness" or "super-uniformity", associated with any parameter. The degree of misfit associated with a parameter estimate can be investigated by comparing the relevant modelled variance, $\Sigma P_{ni}(1-P_{ni})$, with the empirically observed variance, $\Sigma(X_{ni}-P_{ni})^2$.

Two convenient statistics have the form of χ^2 statistics divided by their degrees of freedom. They are the outlier-sensitive, outfit mean-square statistic:

$$U_n = \frac{\sum_i \frac{(X_{ni} - P_{ni})^2}{P_{ni}(1 - P_{ni})}}{\sum_i 1}$$

and the inlier-pattern-sensitive, information weighted, infit mean-square statistic:

$$V_n = \frac{\sum_i (X_{ni} - P_{ni})^2}{\sum_i P_{ni}(1 - P_{ni})}$$

These have proved to be useful indicators of misfit They have expected values of 1. Values less than 1 indicate less local variance in the data than modelled. Values more than 1 indicate excess local variance, noise, in the data. For a typical, minimally constrained analysis the average mean-square value is close to 1.0. Further guidance on the use of these fit statistics to guide response string diagnosis is given in Linacre & Wright (1994).

Model and Real Standard Errors

The highest precision, i.e., minimum standard errors, are obtained when the data fit the Rasch model. Under these circumstances all noise (or lack of noise) in the data is attributed to modelled stochasticity. Then from data $\{X_{ni}\}$, an estimate corresponding to parameter B_n is

$$\hat{B}_n \pm SE(\hat{B}_n)$$

with model standard error given by

$$SE(\hat{B}_n) = \frac{1}{\sqrt{\sum_i P_{ni}(1 - P_{ni})}}$$

When the data fit the model for a parameter, the mean-square fit corresponding to that parameter is 1.0, and the precision of the parameter estimate is the model standard error. As mean-squares depart from 1.0, the data fit the model less exactly, and the overall imprecision of the estimates becomes larger than the model standard errors.

The lowest precision, i.e., maximum standard errors, are obtained when the data misfit the Rasch model as much as possible. Under these circumstances all noise (or lack of noise) in the data is attributed to departures from the model, i.e., unmodelled stochasticity. Then this "real" person measure error variance can be estimated from the Rasch model as

"real" error variance = model error variance *
max (infit mean-square, outfit mean-square, 1.0)

In practice, because the exact amount and source of stochasticity in each observation is unknown, the empirical precision falls at an indeterminable value between the model and real standard errors. In social science, just as in physics, there is a tendency to exaggerate the precision of estimates. "Human beings ... are always inclined to overestimate measuring accuracies" (Albert Einstein). Nevertheless, since there is pressure to report high test reliabilities, the smallest standard errors are usually computed. Still, even the optimistic "model" standard errors are much more meaningful than the self-delusion that the reported measures are point estimates.

Measure Separation

The observed distribution of person measures can be decomposed as:

Observed variance of person measures =
Adjusted (or "true") person measure variance +
person measure error variance

The observed person variance can be computed directly from the reported person measures. The average person measure error variance can be estimated from

$$Error\ variance = \frac{\sum_n SE^2(B_n)}{n}$$

and so the adjusted variance can also be obtained.

A useful indicator of the overall measurement effectiveness of a test is the person separation index. The separation index is

Separation = √ (Adjusted variance / error variance)

The person measure separation index has an value of 1.0 when the variance in the data is divided equally between the variance due to actual differences in person measures and variance due to measurement imprecision. The conventionally good reliability index value of 0.9 corresponds to a separation of 3.0.

Effect of Item Misfit on Person Measurement

Figure 1 shows a typical relationship between test length, person measure standard error and person measure separation for a test containing equally statistically informative items. Each additional item reduces the standard error of person measurement and increases separation.

Figure 2 shows the relationship between stochasticity and separation. If an item is slightly noisy, e.g., if one or two people guess on an MCQ item, then it is still useful for measurement and contributes to the measurement effectiveness of the test. If an item is extremely noisy, e.g., if everyone guesses on an MCQ item, then it is useless for measurement. It increases error variance and lowers person separation and reliability, thus degrading the overall measurement effectiveness of the test.

On the other hand, if an item is slightly muted, e.g., a hard item at the end of a timed test, then it is still useful for measurement and contributes to the measurement effectiveness of the test. If an item is extremely muted, e.g., the accidental inclusion of an item for the second time, then it is useless for measurement. Nevertheless, it appears to decrease overall error variance and so increases person separation and reliability. This is known as the "attenuation paradox".

Empirical-based Investigation of Separation

In order to confirm the effect of misfit on measurement efficiency, the responses of 34 children to the 14 non-extreme items in the Knox Cube Test data (Wright & Stone, 1979, p.33) are used as a core. To this core, several forms of an additional item are added. The following analyses are performed:

1) The original 14 item test.

2) A Guttman item is added. This dichotomous item is as muted as possible. This item has child responses assigned such that high ability children always succeed and low ability children always fail with no stochasticity. This item has mean-square fit less than 1.0.

3) A "whole test" item is added. This rating scale item is as muted as possible. Each response by a child is assigned the rating of the child's score on the original test. Thus, this item duplicates the original test and provides no new information at all.

4) An anti-Guttman item is added. This dichotomous item is as misfitting as possible. This item has child responses assigned such that high ability children always fail and low ability children always succeed with no stochasticity. This item has mean-square fit greater than 1.0.

5) An anti-"whole test" item is added. This rating scale item is as misfitting as possible. Each response by a child is assigned the rating of the child's failure score on the original test. Thus, this item compliments the original test and gives each child the same final total score.

Figure 3 shows the effect on child measure separation when each of these additional items is added, in turn, to the KCT data. In each case, a new Rasch analysis is performed as though the additional item were part of the original data set, i,e., a routine un-anchored analysis. It can be seen that muted, Guttman-like response patterns increase measure separation, but noisy, anti-Guttman patterns decrease separation.

Empirically-based Investigation of Item Effects

Excessively noisy items are clearly detrimental to measurement, and excessively muted items can mislead the analyst as to measurement effectiveness. It is desirable to be able to identify such items. Figure 4 shows the item fit of each additional item under two conditions:

1) An unanchored analysis, "U". This includes an additional item as though it were part of the original analysis. The fit is shown of both the additional item, and of two items, 10 and 12, which are well-fitting with the original data.

2) An anchored analysis, "A". All 14 items and 34 children in the original analysis are anchored at those original measures. The fit of the additional item in this anchored context is shown.

In the original data, both items 10 and 12 have mean-square fits close to 1.0. The addition of Guttman data tends to force these values away from 1.0, while the fit of the Guttman response string itself lies closer to 0.0. Thus, removing Guttman data has the effect of making the rest of the data fit the model better.

The addition of a small amount of noisy data, the anti-Guttman item, makes the reference items, 10 and 12, both appear to slightly overfit the model. The noisy item itself misfits in both the anchored and unanchored runs.

A large amount of noisy data, the anti-whole test item, clearly misfits in the anchored run. In the unanchored run, however, it has the effect of making the whole data set appear random. Thus all person fit statistics now have the value of 1.0. Of course, such anti-Guttman response patterns are easily diagnosed because all child measures collapse to a narrow range, even to a point, and other indicators, such as items fits and point-biserial correlations, become remarkable.

Conclusion

Both very muted and very noisy response strings are seen to be detrimental to measurement. In practice, if such response strings are suspected, analyses with and without such items should be performed to determine the effect of those responses on the overall measurement structure. Noticeably problematic response strings should also be further examined to discover the cause of their doubtful nature. One item may be cuing another or an item may be miskeyed. A response form may have been unintentionally scanned twice or may have been mis-scanned.

Since all empirical response strings contain both Guttman-like and anti-Guttman-like features, excessive diligence at weeding out doubtful items, persons, or individual responses can lead to the elimination of an entire data set! Consequently, measure effectiveness, model fit and utility must be considered together when deciding which subset of the data is most informative for the particular task at hand. Thus one subset of the data may be most informative for item calibration, then another subset for reporting person measures based on the pre-calibrated items.

Linacre JM & Wright BD. 1994. Chi-square fit statistics. Rasch Measurement Transactions 8:2 p. 360-261.

Wright BD & Stone MH. 1979. Best Test Design. MESA Press.

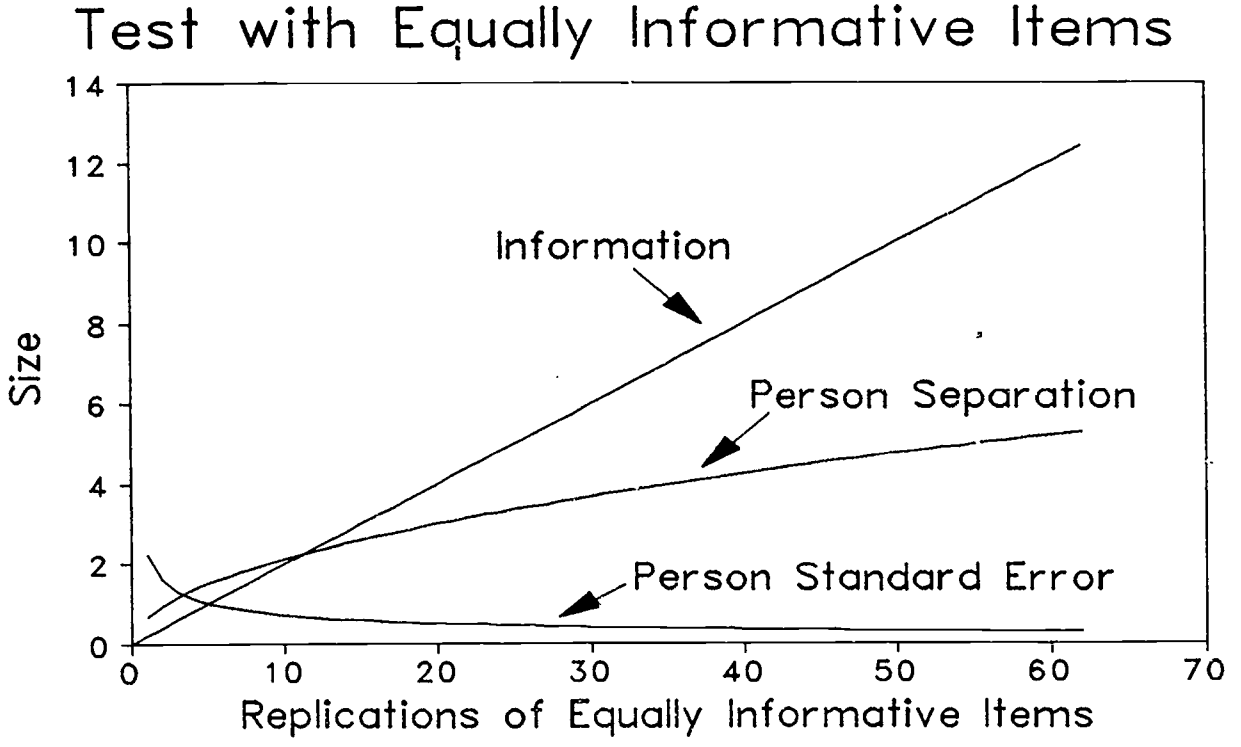

Figure 1 Typical characteristics for test containing equally informative items.

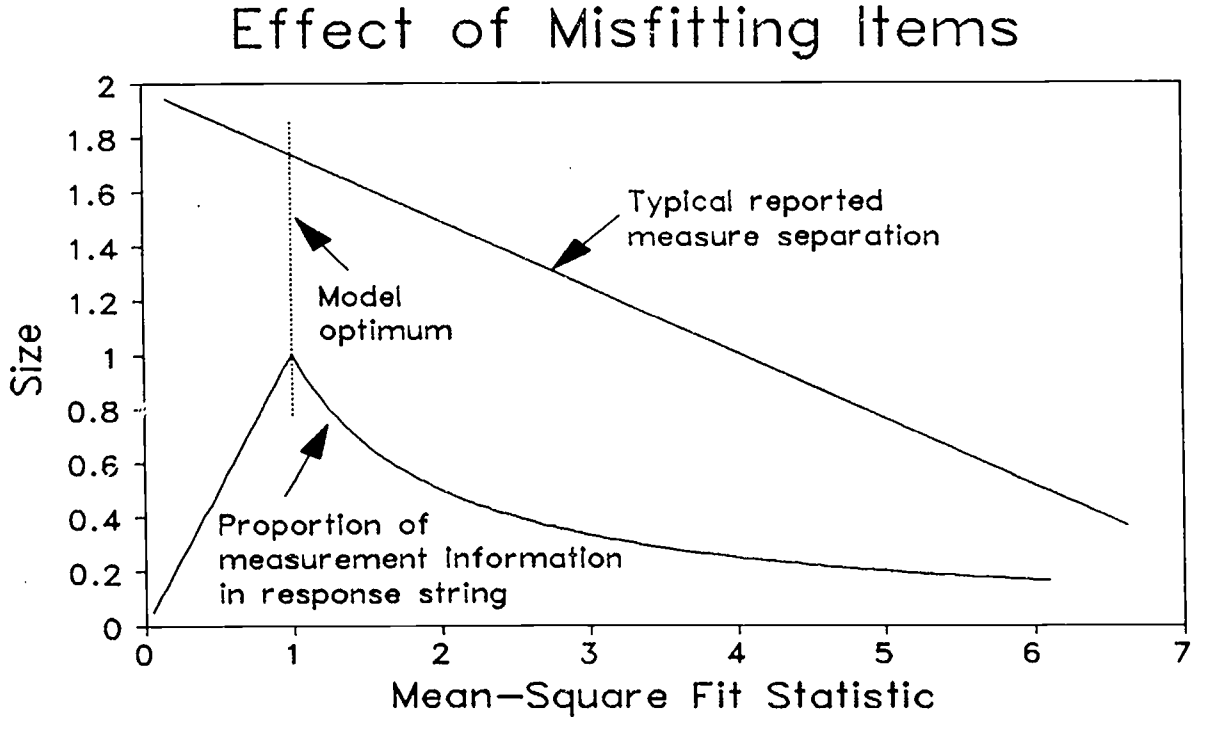

Figure 2 Relationship between mean-square fit and measure separation.

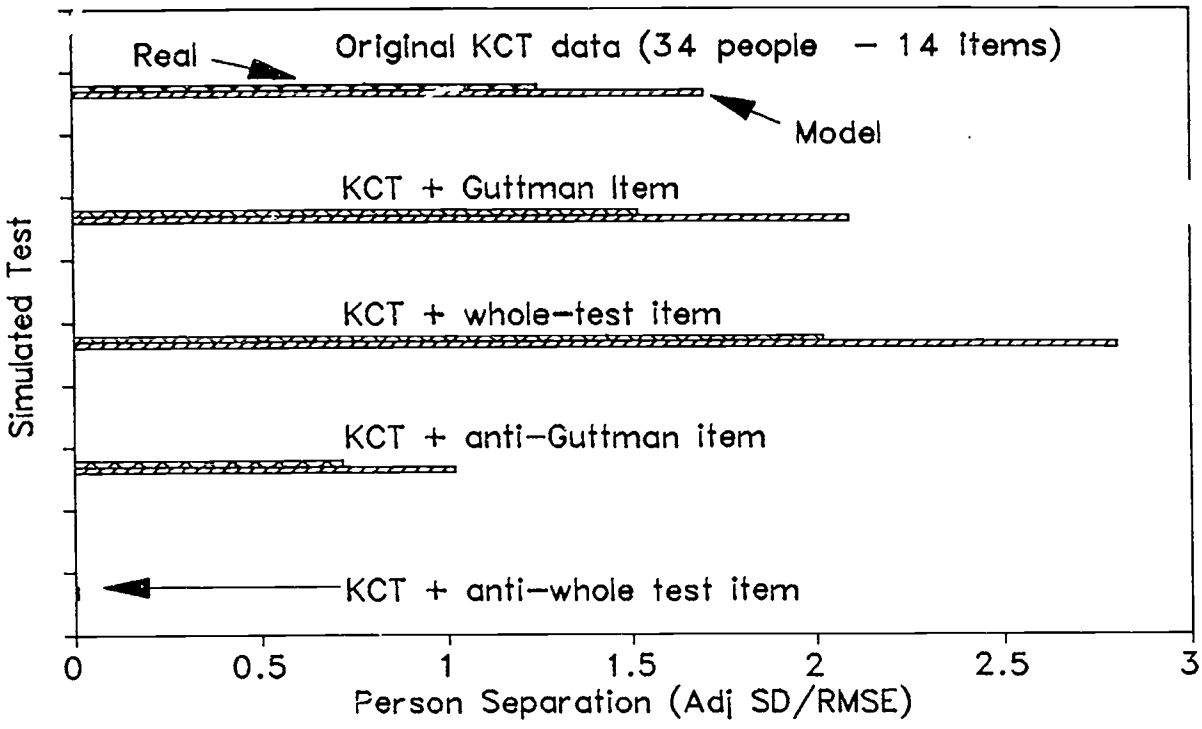

Figure 3 Child separation by KCT data with additional item.

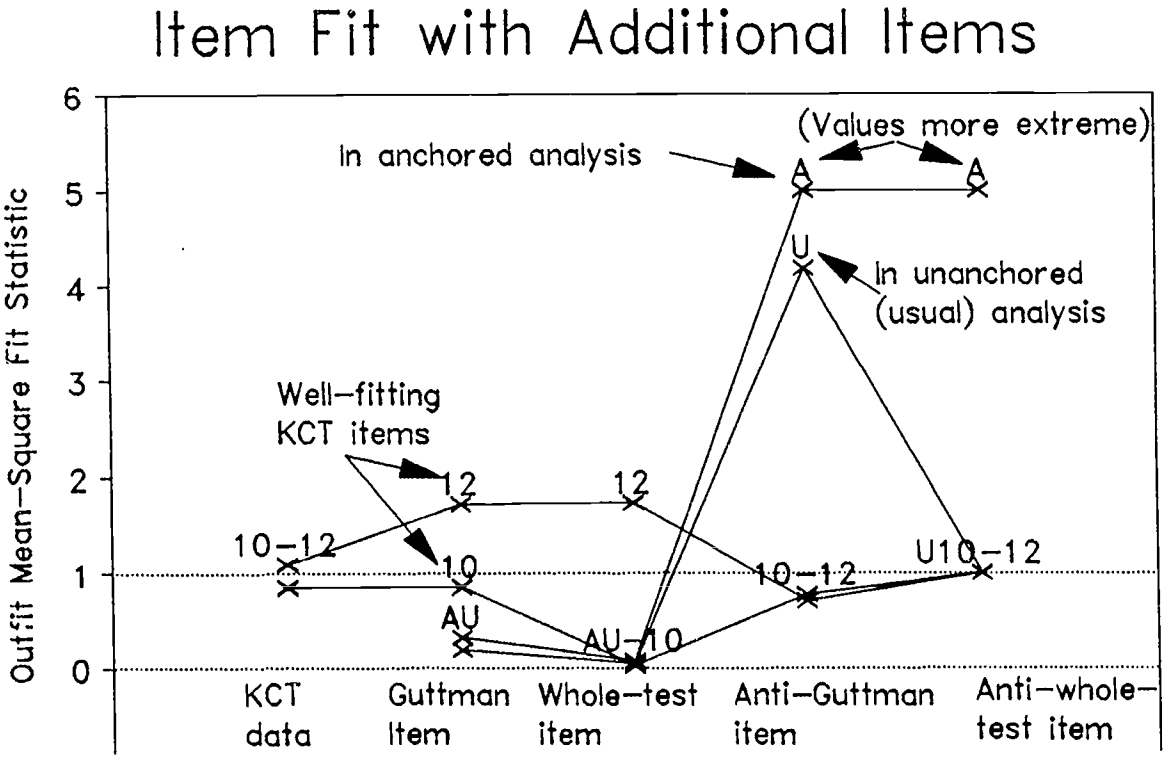

Figure 4 Reported misfit of KCT items.