DOCUMENT RESUME

ED 126 115 TM 005 370

AUTHOR Mead, Ronald

TITLE Assessing the Fit of Data to the Rasch Model.

PUB DATE [Apr 76]

NOTE 15p.; Paper presented at the Annual Meeting of the American Educational Research Association (60th, San Francisco, California, April 19-23, 1976)

EDRS PRICE MF-$0.83 HC-$1.67 Plus Postage.

DESCRIPTORS *Goodness of Fit; Guessing (Tests); *Item Analysis; *Mathematical Models; Response Style (Tests); Statistical Analysis; Test Bias; Test Wiseness

IDENTIFIERS Practice; *Rasch Model; Speededness

ABSTRACT

This paper considers (1) the requirements imposed on data in order to conform to the Rasch model, (2) some common sources of departure from the model, and (3) a procedure for recognizing the occurrence of these disturbances. The specific disturbances discussed are guessing, practice, speededness, and bias. The observed characteristic curve for each situation is compared with the true logistic ogive and with the item characteristic curve that the Rasch estimation procedure would fit to such data. A convenient, interpertable form of residual between model and data is suggested. (Author)

ASSESSING THE FIT OF DATA TO THE RASCH MODEL

Ronald Mead

University of Chicago

Prepared for

American Educational Research Association
Annual Meeting
San Francisco

April, 1976

When Georg Rasch thought about what measurement meant to him, he arrived at the position that in order to obtain something he was willing to call a measurement, the situation had to be dominated by a single person parameter and a single item parameter. This lead him to the mathematical expression:

$$(1)\qquad \text{Prob}(x_{vi}=1|\beta_v, \delta_i)= \frac{e^{\beta_v-\delta_i}}{1+e^{\beta_v-\delta_i}}$$

and the corresponding picture (solid line in Figure 1a) which describes any possible person-item interaction. Contrary to popular belief, this expression does not define a religious cult. It can, however, lead to objective measurement in exchange for some reasonable if sometimes elusive requirements on the situation.

These requirements are:

(a) For a given item, an able person is always more likely to be right than an unable person.

(b) A given person is always more likely to answer an easy item correctly than a difficult one.

This approach to measurement is unique because it is not the result of trying to describe whatever observations happen to look like but rather of deciding what they need to look like to be worthy of the name "measurements." The extent to which Rasch's model can be used to describe the real world has to be investigated empirically. But it is easy to think of cases which do not fit with the model and to understand why they should not be called measurements. Here are several well known examples.

A. Random Guessing (Carelessness)

The model does not allow for random guessing. When this happens, the characteristic curve (Figure 1a) does not approach to zero as the item becomes impossibly difficult. A person of lower ability β_1 has as good a chance of guessing the correct answer as does a person of higher ability β_2.

The same sort of disturbance could occur in the upper tail if very able persons are careless when answering very easy items. Guessing and carelessness will lead us to underestimate the item's difficulty if persons of low ability were ~~used~~ as the calibrating sample and to overestimate its difficulty if persons of high ability are used.

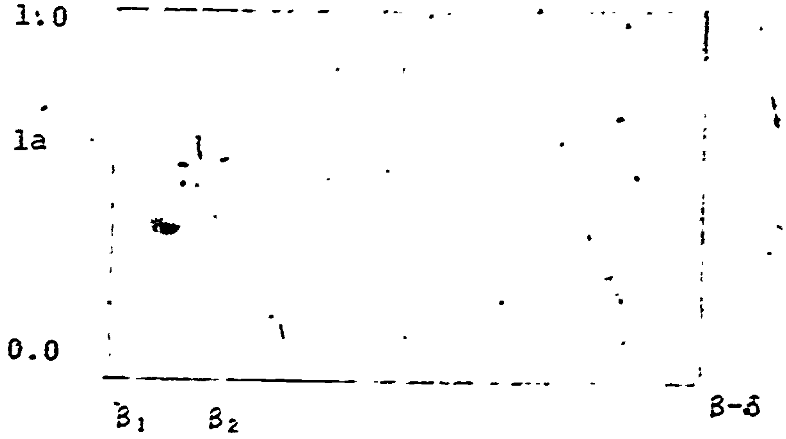


Figure 1a
Guessing &
Carelessness

B. SPEED (PRACTICE)

If people require several items to warm up before they can operate at true ability, the first few items on the instrument will be influenced by lack of practice. Analogously, if people do not finish, the last few items will be influenced by lack of speed. If true ability is unrelated to practice or speed effects, then items affected by practice or speed will seem more difficult and less steep than they are. (Figure 2a). Their outcomes will be influenced by their positions on the instrument. If ability is low enough, there is very little chance of success regardless of the item's position. But for more able people, the likelihood of success is a function of (β-δ) as it should be but also the item's position, combined with how much the person is affected by the position. In the case of practice, after the person has answered a few items, he should be warmed up and able to perform at his true ability. Analogously, for speeded tests, a person's likelihood of success on the last item depends on his probability of success if he attempts the item and his probability of attempting it. These extraneous influences will lessen the item's power to discriminate on the true ability continuum.

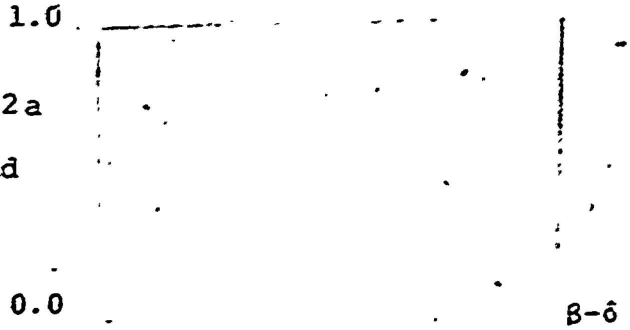


Figure 2a

Practice and Speed

C. BIAS

There is increasing concern for finding items that are fair toward all subpopulations from which we might wish to measure people. One way in which biased items might be expected to operate is, if an item is fair for population A, then the item characteristic curve is of the usual form. However, if the item involves skills or content or behavior beyond the ability of interest, and these skills are harder for a person in population B to acquire than they are for persons in population A, then the item is biased in favor of A. The characteristic curve (Figure 3a) for the item with people from B will be to the left and less steep than the characteristic curve for people from A. It is less steep because the probability of succeeding on the items depends on the probability of having acquired the other behaviors as well as on the position on the ability continuum. For a particular item this shape is the same as that for practice or speed. (In fact, practice and speed can be considered special cases of bias.)

II. Computation of Residuals

The way to determine if these disturbances are present or if it is reasonable to use the model as our explanation for a particular situation is to compare the observed outcome with the predicted outcome.

$$(2) \quad x_{\nu i} = X_{\nu i} - P_{\nu i}$$

When the model does account for the outcome and $P_{\nu i}$ is known, the mean and variance of $x_{\nu i}$ are

$$(3)\quad E(x_{\nu i}) = 0.0$$

$$\mathrm{Var}(x_{\nu i}) = P_{\nu i}(1-P_{\nu i})$$

This residual is the difference between the observed ICC and the predicted ICC in the relative frequency metric, represented by vertical distances on the plots.

Often it is useful to transform to a standard statistic by subtracting its expectation and dividing by its standard deviation:

$$(4)\quad z_{\nu i} = \frac{X_{\nu i}-P_{\nu i}}{\sqrt{P_{\nu i}(1-P_{\nu i})}}$$

which facilitates decisions about the statistical significance of the residual.

We could also look at the difference between the curves in the horizontal direction. The residual in this direction is in ability units (or logits).

Since the derivative of $(\beta_\nu-\delta_i)$ with respect to $P_{\nu i}$ can be viewed as the change in scale from the frequency metric to the ability metric, this can be approximated from $X_{\nu i}-P_{\nu i}$ by

$$(5)\quad y_{\nu i} = x_{\nu i}/(P_{\nu i}(1-P_{\nu i}))$$

with expectation zero and variance one over P times (1-P)

$$(6)\quad E(y_{\nu i}) = 0.0$$

$$\mathrm{Var}(y_{\nu i}) = 1.0/(P_{\nu i}(1-P_{\nu i})) = 1.0/w_{\nu i}$$

Many of the disturbances discussed above have simple relationships to $y_{\nu i}$. which can be conveniently analyzed through weighted least squares using $w_{\nu i} = P_{\nu i}(1-P_{\nu i})$ as the weight.

III. Patterns of Residuals

A. True Difficulties and Abilities are Known

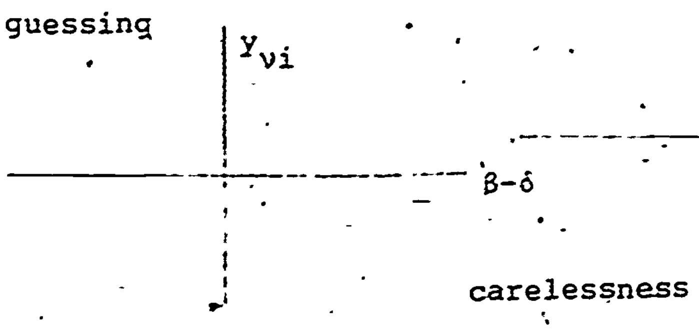
Assuming the true parameters are known, the differences between the model and observed characteristic curves can be easily represented in terms of the residuals. According to the model, the residuals plotted against $(\beta_\nu-\delta_i)$ should fall along a horizontal line through the origin. Disturbances will appear as departures from this horizontal line.

For guessing, the residuals (Figure 1b) follow the horizontal line until the guessing becomes important. Then the residuals are positive since the person is doing better than expected and in that region have a negative trend. For carelessness, the residuals are negative when $(\beta_\nu-\delta_i)$ is large and the slope is again negative.

Figure 1b

Residuals for

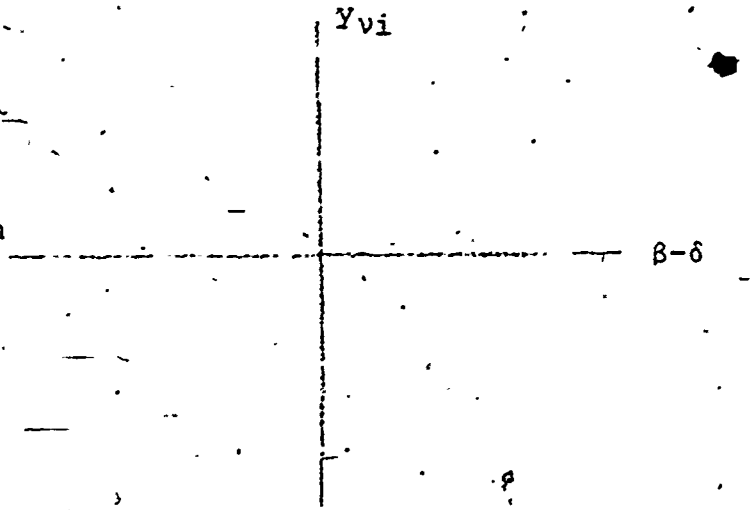
Guessing/Carelessness



If either practice or speed is involved, the items which are affected display negative residuals (Figure 2b) with a negative trend line over the entire range of ability. This must be true since the probability of success if practice or speed is important is less than it should be and the curves become further apart as ability increases. The residual pattern for items biased against a subpopulation is the same shape. The two situations can be distinguished easily since for practice/speed the departures can be organized by item sequence number and for bias, by type of person,

Figure .2b

Residuals For One Item with Practice, Speed or Bias



B. Parameters are not Known

The preceding discussion was greatly simplified by the assumption that the parameters were known. In practice they are not and any disturbances of the sort we have been considering affect our estimates of the parameters. Therefore, the item's characteristic curve that we use for reference in the plots will not be the true ICC but the average of the observed curves, including the distorted ones.

If random guessing is a problem for some items, (Figure 1c) the observed discriminating power for the average ICC will seem to be lowered because of the attempt to fit a simple logistic curve to the guessed upon items. Depending on how the person abilities are distributed we would probably observe misfit over the entire range because of the distortion caused by guessing.

In terms of residuals, we would again observe a pattern (Figure 1d) which might be approximated by a second order polynomial in $(b_\nu-d_i)$. Because the curve is concave upward, it would have a positive quadratic coefficient.

If the problem were carelessness instead, the polynomial would appear concave downward and so have a negative quadratic coefficient. A third order polynomial would be required if both guessing and carelessness were active.

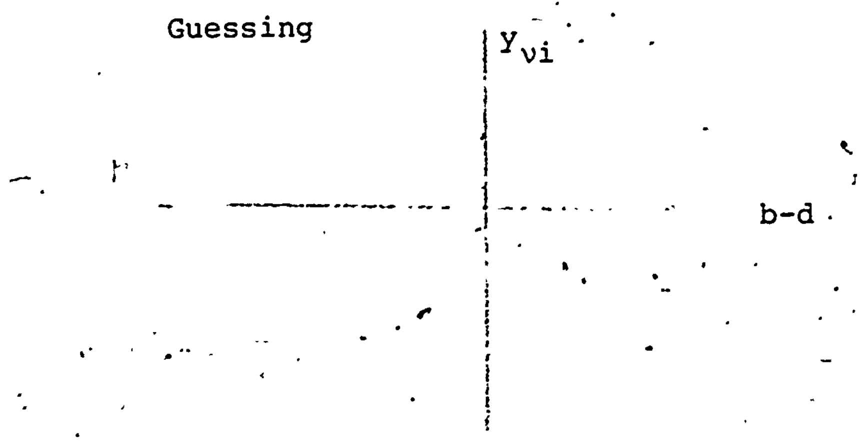


Figure 1a

Residuals For Guessing/Carelessness

The disturbances in the central region, where neither guessing nor carelessness are thought to be operating, are due to the misinformation about ability and difficulty from the actions of guessing and carelessness. The reason significant negative residuals occur with negative values of (b_v-d_i) is because we have overestimated the ability of some persons and underestimated the difficulty of some items. The people were overestimated because they successfully guessed some items and the items were underestimated because some people successfully guessed on them. When we are operating under the assumption that the people are that able and the items are that easy, we are falsely alarmed by some missed items.

Practice and speed, when present, become an unwanted part of the "variable" that we observe. Practiced people score higher than unpracticed people of the same ability by doing well on early items. Fast people answer the last items and so score higher than slow people of the same ability. These phenomona would give the items affected the appearance of high discrimination which shows up as a positive trend in the residuals with respect to (b_v-d_i). But this apparent discrimination is with respect to the "variable" that includes practice and speed effects, not the variable that we set out to measure.

Biased items could produce apparent high discriminations in a similar way. Persons in the favored group would tend to score high on the entire instrument but particularly on the most unfair items. These items would appear to have high discriminations but the inflation in their discriminations would be due to their power to distinguish between the favored and unfavored groups rather than between more able and less able persons. (Figures 3c and d).

If we instead plot the residuals for persons against b-d, we usually find that the residual have negative trends for persons from ~~the unfavored~~ group and positive trends for persons in the favored group. (Figure 3e) This is because for the unfavored group, the items are "measuring" two variables and so have relatively less information about the variable that is of interest to us. This means that even if all items are unfair, the presence of a general bias should be indicated by a decrease in item discriminating power.

It must be kept in mind that when dealing with departures from the Rasch model, we have not achieved sample-free calibration. The patterns in the residual plots depend very much on who was in the calibration sample. The appearance of these effects will depend on the relative numbers ~~of fair~~ versus unfair items and favored versus unfavored persons.

Conclusion

I began this discussion by asking what sorts of things would cause data not to fit the Rasch model. After listing a few common disturbances, I discussed the straightforward approach that Prof. Wright and I have been using to discover if any of these disturbances are present. While our approach incorporates the use of well known least squares techniques, much work remains to be done before the properties of these techniques are well understood in our application.

All of the disturbances considered represent some form of multidimensionality; they would violate any model that assumes unidimensionality. Since the effect of the disturbances often appears as a change in the slope of the ICC, any model which includes item discrimination as a parameter would appear to fit such data. Thus we would be in the unfortunate situation of accepting data which violate the model's assumptions but pass the tests of fit.

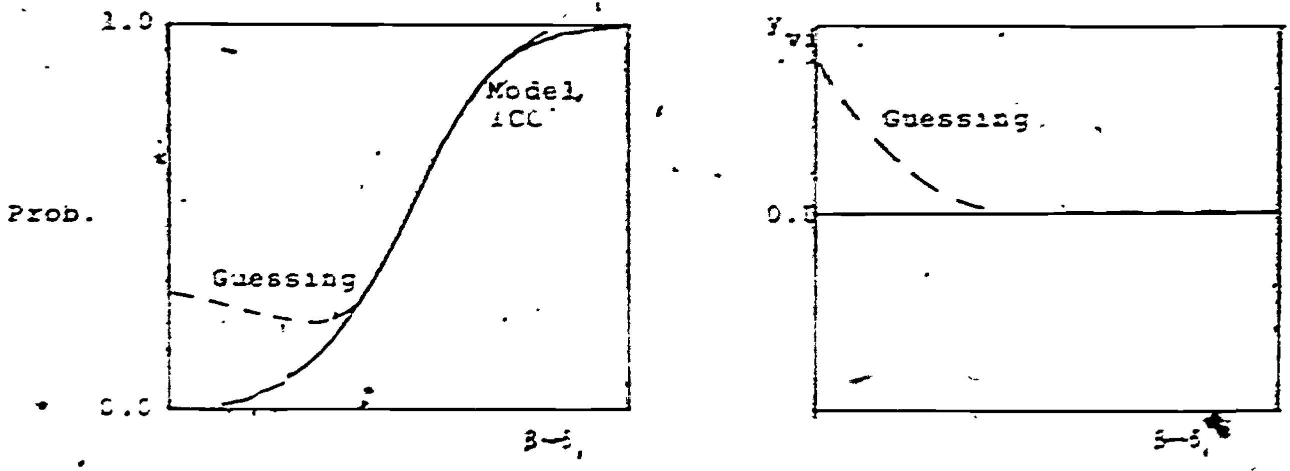
By fitting such a general model, we would not only have lost the desireable measurement properties of the Rasch model, but we would have also mislead ourselves about the true nature of the variable that we were seeking. When we understand a process well enough to control its multidimensionality, we have no need for any additional parameters. Until then, I do not think we can afford them.

ITEM CHARACTERISTIC CURVES

ITEM RESIDUALS

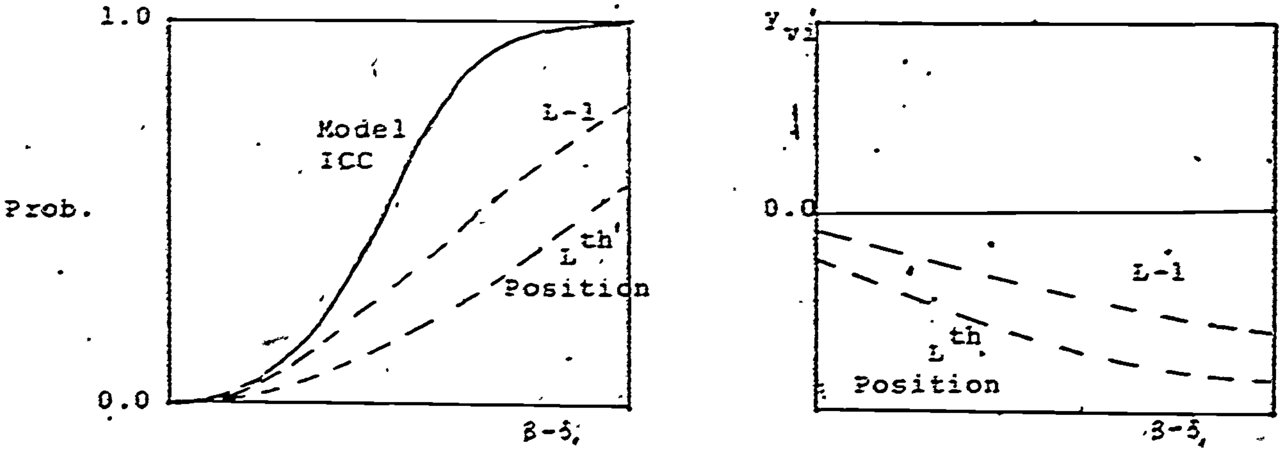
Figure 1a: Guessing

1b.



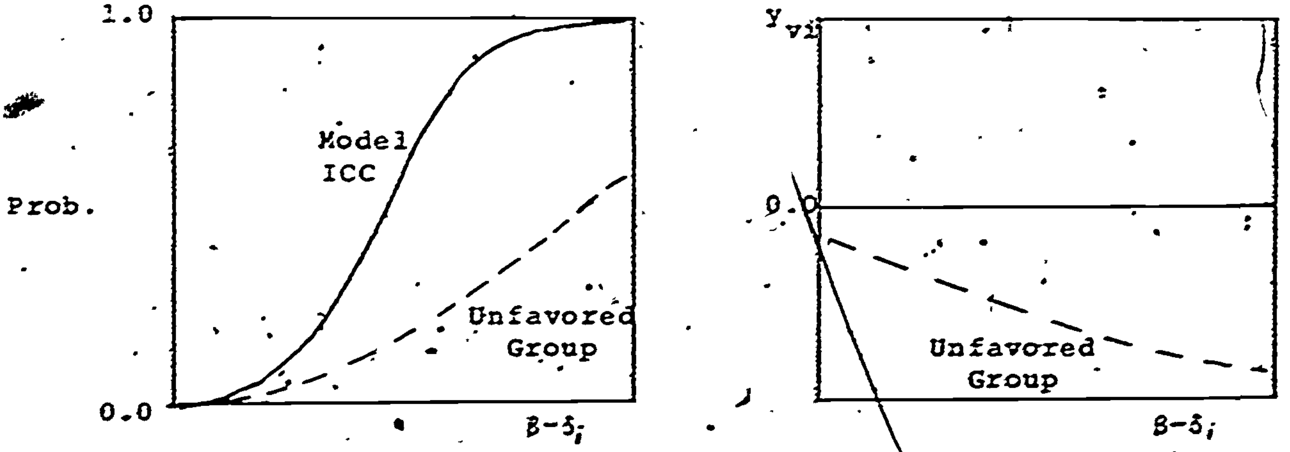
2a: Speed (or Practice)

2b.



3a: Bias

3b.



Mead and Wright
April, 1976

Observed and Fitted Characteristic Curves

Figure 1c: Guessing ICC's

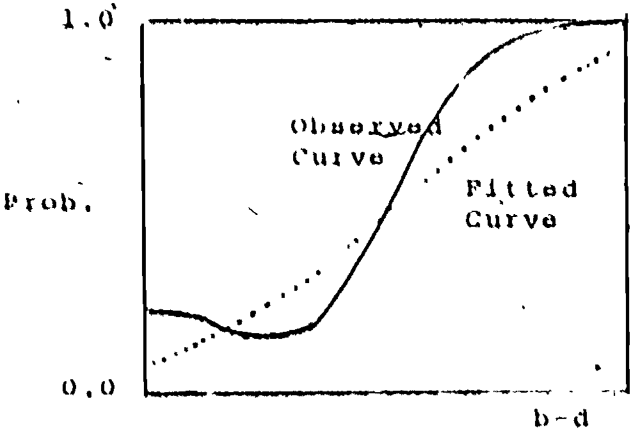
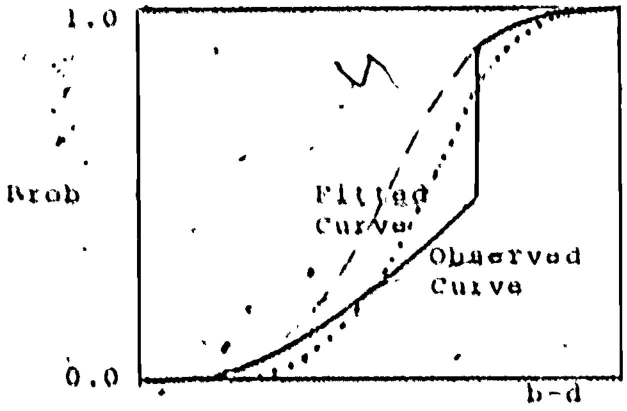


Figure 3c: Bias ICC's



Item Residuals

Figure 1d: Guessing Residual For Item k

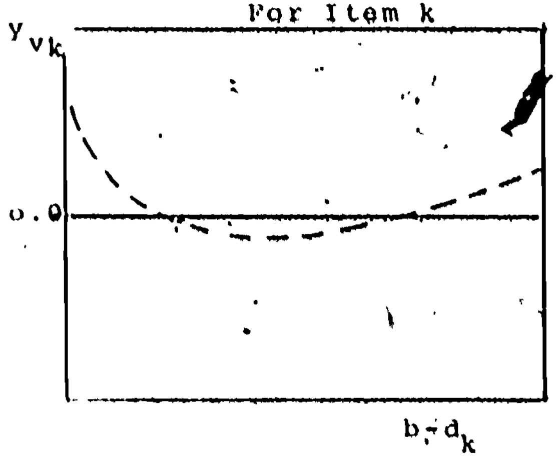
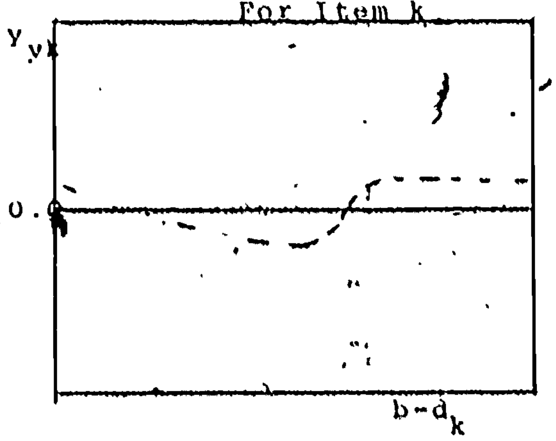


Figure 3d: Bias Residuals For Item k



Person Residuals

Figure 1e: Guessing Reiduals For Person u

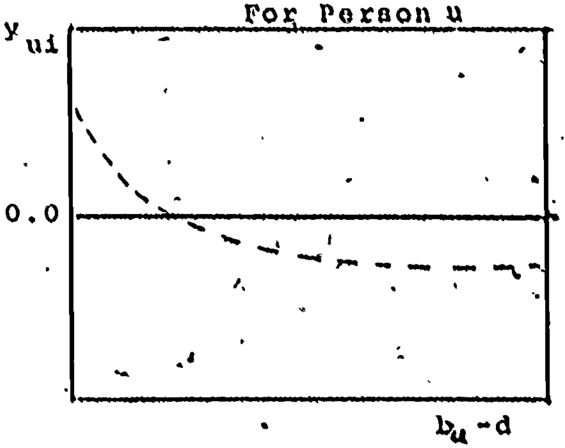
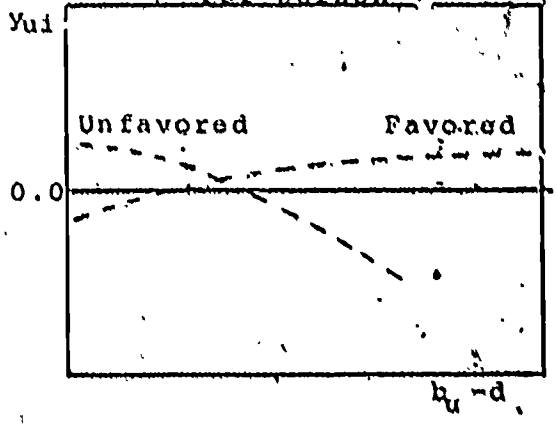


Figure 3e: Bias residuals For person u



Mead and Wright
April, 1976