7. FIT ANALYSIS

The Rasch model specifies the relation that must dominate what happens when a person takes an item for the resulting responses to be useful for measurement. A complete analysis must include an evaluation of how well the data fit this essential specification. If a person answers the hard items on a test correctly but misses several easy items, we are surprised by the resulting implausible pattern of incorrect responses. While we could examine individual records item-by-item to determine this kind of invalidity, in practice we want to put such evaluations on a systematic and manageable basis. We want to be specific but also objective in our reaction to implausible and hence invalid observations.

Even when a particular application tends to fit the measurement model, we cannot predict in advance how well new items or old ones will continue to work in every new situation to which they might be applied. We cannot know in advance how all persons will always respond. Therefore, if we are serious in our intention to measure, we must examine every application to see how well each new set of responses corresponds to our measurement intentions. We must evaluate not only the plausibility of the sample of persons' responses, but also the plausibility of each persons' responses to their set of items. To do this we must examine the response of each person to each item to determine whether that response is consistent with the general pattern of responses observed.

We begin fit analysis by examining the data resulting from the administration of a test of L items to a sample of N persons producing an N x L matrix of responses with every row consisting of the responses of each person n to the L items and every column consisting of the responses of the N persons to each item i. When the responses are dichotomous, the resulting matrix will consist of correct $(X_{ni} = 1)$ and incorrect $(X_{ni} = 0)$ responses.

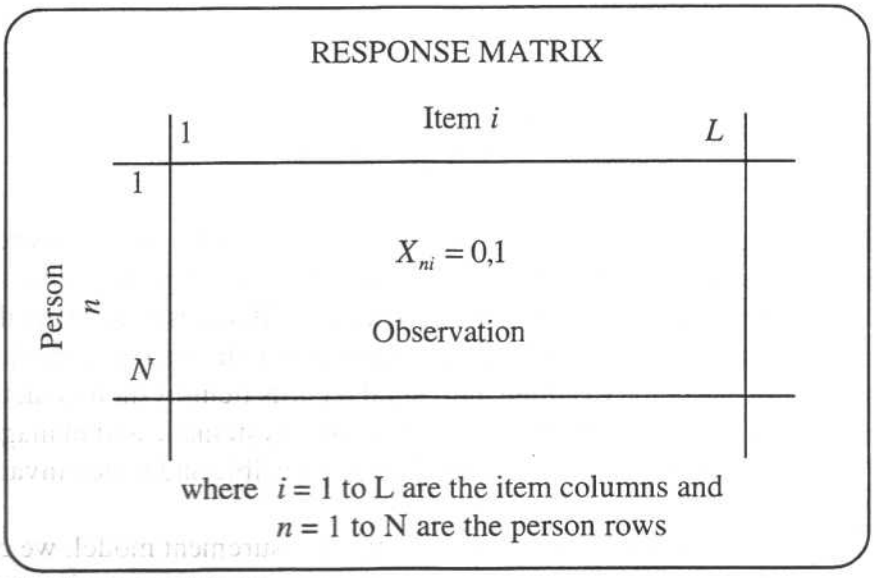
The construction of useful measures and calibrations does not require that these data be complete. The particular items addressed by each person can vary as long as there is a sufficient network of overlaps to connect the entire matrix. For simplicity here, however, we will carry out the explanation as though data were complete.

From the N x L data matrix of $X_{ni} = 0$ or 1 we count the item scores S_i and person scores R_n used to estimate the abilities of persons B_n and the difficulties of items D_i. Procedures for this are explained in Best Test Design (Wright and Stone, 1979, pp. 28-65).

This chapter explains the analysis of fit (Wright and Stone, 1979, pp. 66-82 and 165-181).

RESIDUAL FROM EXPECTATION

To evaluate fit we compare the observed person and item responses X_{ni} to the expected values P_{ni} that are determined for them by the measurement model. The expected value of the dichotomous observation X_{ni} is $P_{ni} = \exp(B_n - D_i) / [1 + \exp(B_n - D_i)]$.



where $i = 1$ to L are the item columns and
$n = 1$ to N are the person rows

A consequence of the Rasch model is that the person right answer count, a total score for an individual, contains all of the information needed to measure that person and the item right answer count, a total score for an item, contains all the information needed to estimate the difficulty of that item. That is to say, that right answer counts are sufficient statistics for estimating person measures and item calibrations.

RASCH MODEL EXPECTATIONS

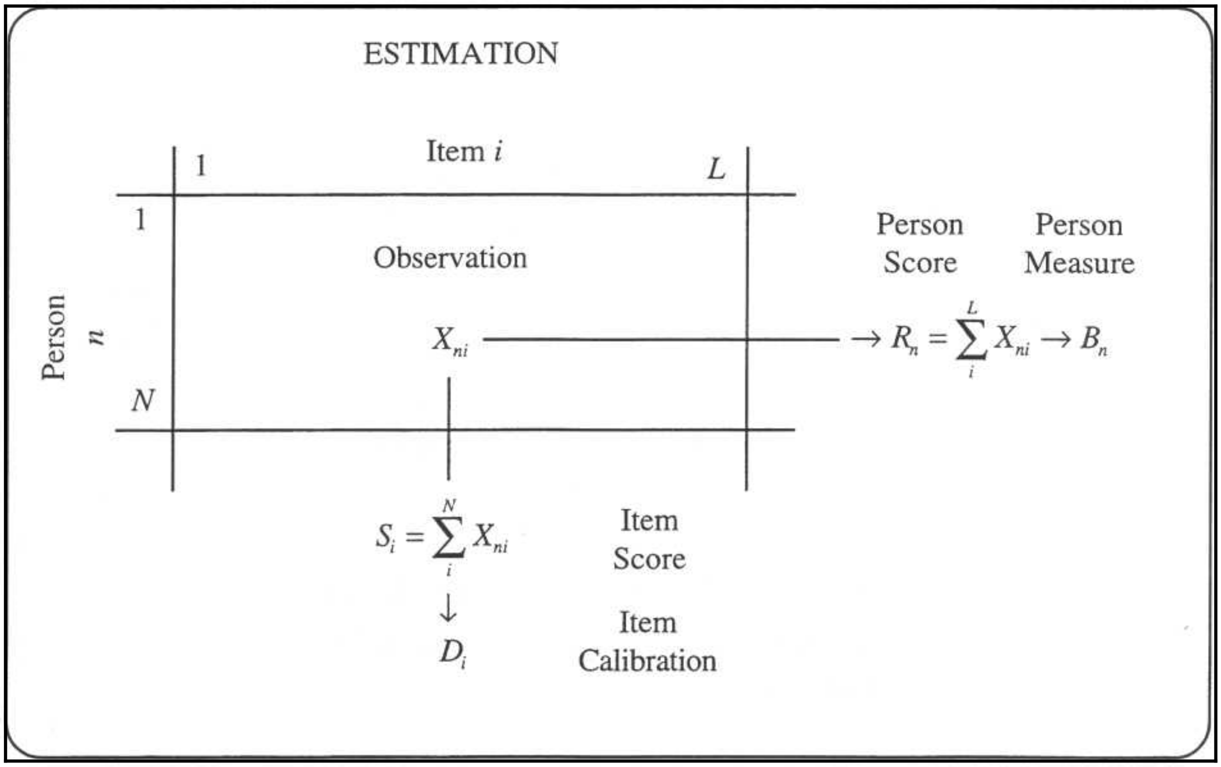
The Rasch model is derived from the requirement that person measures and item calibrations be separately estimable. This requires that (1) a more able person always have a greater probability of success on any item than a less able person, and (2) any person always be more likely to do better on an easier item than on a harder one. Fit analysis evaluates the extent to which particular data serve this fundamental requirement for measurement. Fit analysis shows us the extent to which any data can be used to construct measures. Each data analysis must include an evaluation of how well those particular data fit the expectations which measurement requires.

When an observed pattern of responses shows significant deviation from measurement expectations, we can use the particulars of the measurement model together with the person and item estimates to calculate a statistical index of unexpectedness for any particular response or any subset of responses including all of the responses to a particular item or all of the responses made by a particular person.

DETERMINING FIT

The procedure for analysis of fit involves the three steps:

1. Responses	2. Model	3. Residuals
i	i	i
n $\boxed{X_{ni} = 0,\ 1}$	n $\boxed{0 < P_{ni} < 1}$	n $\boxed{Y_{ni} = X_{ni} - P_{ni}}$
What we observe	What we expect	The difference



R_n is the sum $\sum_i^L$ of the person responses X_{ni} over item $i = 1, L$.

B_n is the person measure estimated from R_n .

S_i is the sum $\sum_n^N$ of the item responses X_{ni} over persons $n = 1, N$.

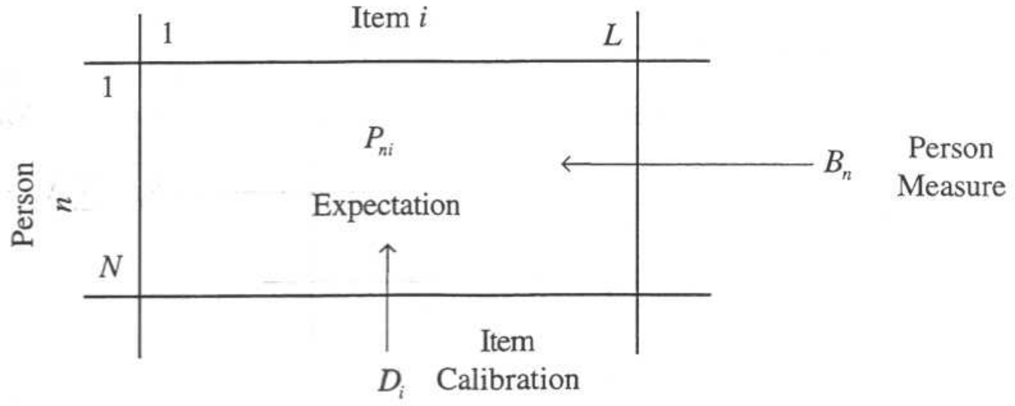
D_i is the item difficulty estimated from S_i .

EXPECTATIONS

To observe response plausibility, validity or fit we calculate the difference $(B_n - D_i)$ between the estimates of person ability B_n and item difficulty D_i for each person n and item i. When this difference is positive it means that the item should be easy for the person. The more positive this difference, the easier the item is expected to be and hence the greater our expectation that the person will succeed on that item and $X_{ni} = 1$.

When the difference is negative, however, the item should be difficult for the person. The more negative the difference $(B_n - D_i)$ becomes, the more difficult the item should be for the person and hence the greater our expectation that the person will fail on that item and $X_{ni} = 0$.

PREDICTION



$P_{ni} = \exp(B_n - D_i) / [1 + \exp(B_n - D_i)]$ and B_n and D_i are the person measure and item calibration for person n and item i.

Chi-square and mean square goodness-of-fit statistics can be constructed from the residual difference $Y_{ni} = X_{ni} - P_{ni}$ between the observed X_{ni} and its expectation P_{ni}. This residual quantifies the fit between each person n and each item i.

The model estimates the expected value or probability of dichotomous response $X_{ni} = 1$ as:

$$P_{ni} \exp(B_n - D_i) / [1 + \exp(B_n - D_i)]$$

where B_n = the estimated ability measure of person n

D_i = the estimated difficulty calibration of item i

and P_{ni} = the probability that $X_{ni} = 1$.

RESIDUALS

The probability P_{ni} is an estimate of the expected value of instances of the response $X_{ni} = 0,\ 1$. The expected binomial variance of these instances of X_{ni} around P_{ni} is estimated by $Q_{ni} = P_{ni}(1 - P_{ni})$. These expectations P_{ni} and variances Q_{ni} can be combined to form a standardized residual Z_{ni} for each X_{ni}: $Z_{ni} = (X_{ni} - P_{ni}) / [P_{ni}(1 - P_{ni})] = (X_{ni} - P_{ni}) / Q_{ni}^{1/2}$.

We estimate this standardized residual Z_{ni} by subtracting from the observed X_{ni} its estimated expected value P_{ni} and dividing the difference by its expected standard deviation $Q_{ni}^{1/2} = [P_{ni}(1 - P_{ni})]^{1/2}$.

This standardized residual Z_{ni} has a logistic distribution with an expected mean of 0 and a variance of 1. The reference values of 0 and 1 help us to evaluate the extent to which the standardized residuals deviate from their model expectations.

Examination of residuals shows us whether we can proceed to use the items to make valid measures or whether further work on the items are required in order to bring the testing elements into line with the intended plan. Examination of person residuals indicates the extent to which persons have responded to the test in the expected manner. Since X_{ni} takes only the two values of "0" and "1", the two values for the standardized residuals can be expressed in terms of the estimates B_n and D_i and the observed response X_{ni} .

Thus, $Z_{ni}^2 = [\exp(B_n - D_i)]$ can be used to indicate the unexpectedness of an incorrect response X_{ni} 0 to a relatively easy item, while $Z_{ni}^2 = [\exp(D_i - B_n)]$ can be used to indicate the unexpectedness of a correct response $X_{ni} = 1$ to a relatively hard item. These two expressions can be combined into one as $Z_{ni}^2 = \exp[(2X_{ni} - 1)(D_i - B_n)]$.

The values of this Z_{ni}^2 can be ascertained for each X_{ni} of 0 or 1 and then accumulated over items to evaluate the plausibility of any person measure, or over persons to evaluate the plausibility of any item calibration.

To evaluate any unexpected response X_{ni} we quantify its unexpectedness from the difference between the ability measure of that person B_n and the difficulty calibration for that item D_i. For example, an unexpected incorrect response of $X_{ni} = 0$ associated with a person ability $B_n = -1.2$ and an item difficulty of $D_i = -3.9$ produces a difference $(B_n - D_i)$ of $[(-1.2) - (-3.9)] = +2.7$ and a squared standard residual of $Z_{ni}^2 = \exp(2.7) = 14.9$.

We associate unexpected incorrect answers $X_{ni} = 0$ with $(B_n - D_i)$ and unexpected correct answers $X_{ni} = 1$ with $(D_i - B_n)$ because when the response is incorrect, and $X_{ni} = 0$, then the index of unexpectedness is $Z_{ni}^2 = \exp(B_n - D_i)$, but when the response is correct, $X_{ni} = 1$, then the index is $Z_{ni}^2 = \exp(D_i - B_n)$.

Unexpectedness is always marked by a positive difference, either $(B_n - D_i)$ or $(D_i - B_n)$. The values for Z_{ni}^2 can be looked up in Table 7.1 which gives either values of $Z_{ni}^2 = \exp(B_n - D_i)$ for unexpected incorrect answers $X_{ni} = 0$ or values of $Z_{ni}^2 = \exp(D_i - B_n)$ for unexpected correct answers $X_{ni} = 1$.

Thus, the entry C_x in Column 1 of Table 7.1 is either $C_0 = (B_n - D_i)$ when $X_{ni} = 0$ and the response is incorrect or $C_1 = (D_i - B_n)$ when $X_{ni} = 1$ and the response is correct.

Column 3 of Table 7.1 gives the improbability of the observed response $P_{ni} = 1/(1 + Z_{ni}^2)$.

Table 7.1

Evaluating Unexpectedness

1. Difference Between Person Ability and Item Difficulty C_x	2. Squared Standardized Residual $Z^2 = \exp\ C_x$	3. Improbability of the Response $P = 1/(1+Z^2)$
-0.6, 0.4	1	.50
0.5, 0.9	2	.33
1.0, 1.2	3	.25
1.3, 1.5	4	.20
1.6, 1.7	5	.17
1.8, 1.8	6	.14
1.9, 2.0	7	.12
2.1	8	.11
2.2	9	.10
2.3	10	.09
2.4	11	.08
2.5	12	.08
2.6	13	.07
2.7	15	.06
2.8	16	.06
2.9	18	.05
3.0	20	.05
3.1	22	.04
3.2	25	.04
3.3	27	.04
3.4	30	.03
3.5	33	.03
3.6	37	.03
3.7	40	.02
3.8	45	.02
3.9	49	.02
4.0	55	.02
4.1	60	.02
4.2	67	.02
4.3	74	.01
4.4	81	.01
4.5	90	.01
4.6	99	.01

This improbability P_{ni} provides a significance level for the null hypothesis of fit for any particular response. With our example of $(B_n - D_i) = 2.7$ we have a significance level of $P_{ni} = .06$ against the null hypothesis that the response of the person to this item is consistent with the model.

When the Z_{ni}^2 are accumulated over items for a person or over persons for an item, simulations have shown that the resulting sums can be usefully evaluated by chi-square distributions with L - 1 degrees of freedom for a person and N - 1 degrees of freedom for an item.

These fit statistics are called "outfits" because they are heavily influenced by outlying, off-target, unexpected responses. A useful alternative is to weigh residuals by the information they contain so that the fit statistics are information weighted or "infits" and hence focus on inlying, on-target, unexpected responses. The calculations for each type of fit statistic are outlined in the summary section.

SUMMARY

The following formulas summarize the calculations for the analysis of dichotomous fit.

Observed Response: $X_{ni} = 0,\ 1$

Expected Response: $P_{ni} = \exp(b_n - d_i) / \left[1 + \exp(b_n - d_i)\right]$

Response Variance: $Q_{ni} = P_{ni}(1 - P_{ni})$

Score Residual: $Y_{ni} = X_{ni} - P_{ni}$

Standardized Residual: $Z_{ni} = Y_{ni} / Q_{ni}^{1/2}$

Fit Mean Square:

Outfit: $U_n = \sum_{i}^{L} Z_{ni}^2 / L$

$U_i = \sum_{n}^{N} Z_{ni}^2 / N$

Infit: $V_n = \sum_{i}^{L} Y_{ni}^2 / \sum_{i}^{L} Q_{ni}$

$V_i = \sum_{n}^{N} Y_{ni}^2 / \sum_{n}^{N} Q_{ni}$

Fit Standard Errors:

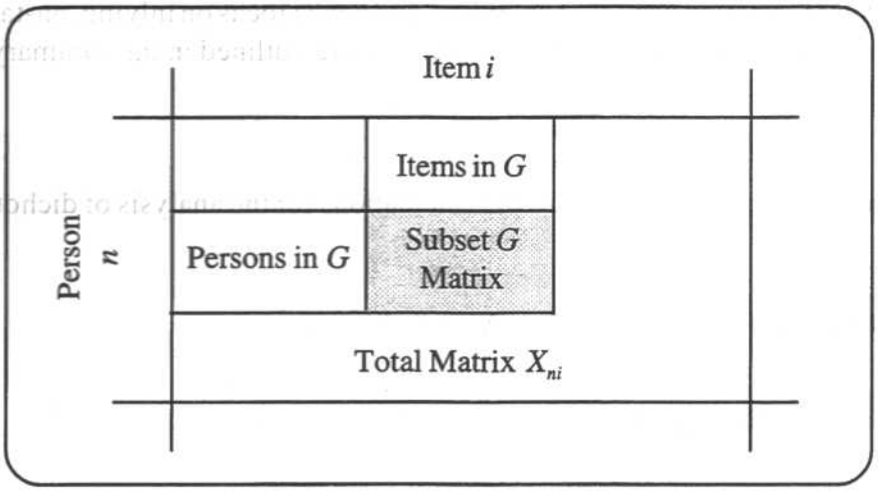
Outfit: $SE_u = \left[\Sigma(1/Q - 4)\right]^{1/2} / \Sigma 1$

Infit: $SE_v = \left(\sum Q - 4\sum Q^2\right)^{1/2} / \sum Q$

Fit Standardization: $T_u = \left(U^{1/3} - 1\right)\left(3 / SE_u\right) + \left(SE_u / 3\right)$

$$T_v = \left(V^{1/3} - 1\right)\left(3 / SE_v\right) + \left(SE_v / 3\right)$$

Fit analysis can also be done for subsets of person-item responses taken from the total matrix of responses. In this manner we can evaluate the responses of any person or subset of persons to any item or subset of items or evaluate any item or subset of items by any person or subset of persons:



For the analysis of any subset (G) of the data matrix of X_{ni} use the following formulas:

Logit Bias: $G = \sum^{G} Y_{ni} / \sum^{G} Q_{ni}$

Standard Error: $SE_G = \left(\sum^{G} Q_{ni}\right)^{-1/2} = 1 / \left(\sum^{G} Q_{ni}\right)^{1/2}$

Mean Score Residual in G: $Y_G = \sum^{G} Y_{ni} / \sum^{G} 1$

Infit Noise in G: $V_G = \sum^{G} \left(Y_{ni} - Y_G\right)^2 / \sum^{G} Q_{ni}$

Standard Error: $SEV_G = \left(\sum^{G} Q_{ni} - 4\sum^{G} Q_{ni}^2\right)^{1/2} / \sum^{G} Q_{ni}$

where $\sum^{G}$ means summed over n and i in G.

The analysis of fit evaluates how well our data cooperate with the construction of measurement. Analysis of fit gives us a tool to monitor the responses of persons and items. We can evaluate any set of items or persons to determine where misfit occurs. Fit analysis provides the quality control technique required to supervise and validate test items and person responses. When fit is within our guidelines, we have the control required to feel confident about item calibration and person measurement. When misfit is discovered we can locate its occasions and begin further study of the items or persons involved.

The analysis of fit is never completed because continued use of the instrument requires that we constantly monitor item and person responses to maintain quality control.

MEASUREMENT ESSENTIALS

2nd Edition

BENJAMIN WRIGHT

MARK STONE

WIDE RANGE, INC.
Wilmington, Delaware