

16. PARAMETER ESTIMATION

Parameter estimation for Rasch measurement is usually done by a computer program like BICAL (Wright & Mead, 1976), BIGSCALE (Wright, Linacre & Schulz, 1990) or BIGSTEPS (Wright & Linacre, 1997).

The PROX procedure, however, is a method of estimation so easy to apply that it is completely manageable by hand. The simplicity of PROX is useful because it details exactly how the Rasch model works in practice. PROX accomplishes the primary aims of Rasch item analysis:

- 1) linearization of item raw scores (P-values) onto an interval scale with relevant errors of calibrations and
- 2) adjustment for the sampling effects of person ability.

In so doing, PROX almost always approximates the results obtained by more elaborate procedures extremely well.

The simplification which enables PROX is to approximate the effects on item calibration of sample ability with a sample mean and standard deviation and the effects on person measurement of test item difficulty with a test mean and standard deviation. This simplification makes PROX easy to apply by hand. Nothing more than the observed distributions of item and persons scores, a hand calculator and pencil and paper are needed.

PROX is as applicable to large assessment problems like national item banking as it is to evaluating a small classroom of examinees. The PROX algorithm is the working basis of most successful computer assisted testing (CAT) programs.

In this chapter we

- 1) outline the PROX equations, and
- 2) explain how these equations implement Rasch measurement.

The chapter can be used to guide the user in calibrating items and measuring persons from their own data. An example worked out in numerical detail can be found in the second chapter of *Best Test Design* (Wright & Stone, 1979, pp. 28-45).

THE PROX ESTIMATION EQUATIONS

PROX simplifies the representation of person abilities b_n to a normal distribution with sample mean ability M and sample ability standard deviation σ and the representation of item difficulties d_i to a normal distribution with test item mean difficulty H and test item difficulty standard deviation w .

When that is done, then the measure b_n for person n with person score r_n on a test of L items becomes

$$b_n = H + X \log[r_n / (L - r_n)] \quad 16.1$$

where

H = mean difficulty of the L items taken,
 X = the scaling necessary to adjust for the
difficulty standard deviation w of these L
items, r = the raw test score of person n ,

and the calibration d_i for item i with item score s_i from a sample of N persons is

$$d_i = M + Y \log[(N - s_i) / s_i] \quad 16.2$$

where

$d_i = M + Y \log[(N - s_i) / s_i]$
 M = mean ability of the N persons taking the test,
 Y = the scaling necessary to adjust for the ability
standard deviation σ of these N persons,
 s_i = the raw sample score of item i .

and

$$X = [1 + (w^2 / 2.89)]^{1/2} \cong [1 + (w^2 / 5.8)] \quad 16.3$$

$$Y = [1 + (\sigma^2 / 2.89)]^{1/2} \cong [1 + (\sigma^2 / 5.8)]. \quad 16.4$$

The divisor $2.89 = 1.7^2$ comes from the scaling factor 1.7 which, because the logistic ogive for values of $1.7z$, is never more than one percent different from the normal ogive for values of z , brings the cumulative logistic distribution into approximate coincidence with the cumulative normal distribution. PROX uses this coincidence to obtain its simplification.

The estimates b_n and d_i have standard errors

$$SE(b_n) = X[L / r_n(L - r_n)]^{1/2} \cong 2.5 / L^{1/2} \quad 16.5$$

$$SE(d_i) = Y[N / s_i(N - s_i)]^{1/2} \cong 2.5 / N^{1/2} . \quad 16.6$$

APPLYING THE PROX ESTIMATION EQUATIONS

This estimation method can be applied to observed item scores s_i by calculating the sample score logit of item i as

$$x_i = \log[(N - s_i) / s_i] \quad 16.7$$

and to the observed person scores r_n by calculating the test score logit of person n as

$$y_n = \log[r_n / (L - r_n)] \quad 16.8$$

The scaling coefficients X and Y can be estimated from

$$X = \{[1 + (U / 2.89)] / [1 - (UV / 8.35)]\} \quad 16.9$$

for the person logit scaling coefficient and

$$Y = \{[1 + (V / 2.89)] / [1 - (UV / 8.35)]\} \quad 16.10$$

for the item logit scaling coefficient.

Where $U = (\sum_i x_i^2 - Lx_*^2) / (L-1)$ and $8.35 = 2.89^2 = 1.7^4$,

$$U = \left(\sum_i x_i^2 - Lx_*^2 \right) / (L-1) \quad 16.11$$

is the item logit variance,

$$V = \left(\sum_n y_n^2 - Ny_*^2 \right) / (N-1) \quad 16.12$$

is the person logit variance, $x_* = \sum_i x_i / L$ is the item logit mean, and $y_* = \sum_n y_n / N$ is the person logit mean.

To complete the estimation, we anchor the scale "zero" at the center of the test by defining $H \equiv 0$ so that

$$d_i = N| + Yx_i = Y(x_i - x_*) \text{ because } M = -Yx_* \quad 16.13$$

for each item difficulty and

$$b_n = H + Xy_n = Xy_n \text{ because } H \equiv 0 \quad 16.14$$

for each person ability.

STANDARD ERRORS

The standard errors of these person and item estimates are

$$SE(b_n) = X[L / r_n(L - r_n)]^{1/2} \cong 2.5 / L^{1/2} \quad 16.15$$

and

$$SE(d_i) = Y[N / s_i(N - s_i)]^{1/2} \cong 2.5 / N^{1/2} \quad 16.16$$

SAMPLE STATISTICS

The estimates of the person sample mean M and standard deviation σ are

$$M \approx -Yx. \quad 16.17$$

$$\sigma \approx 1.7(Y^2 - 1)^{1/2} \quad 16.18$$

ANALYSIS OF RESIDUALS

When we have estimated b_n and w we can use them to obtain the difference between the model's prediction and the data observed. Residuals from the model are calculated by estimating from b_n and d_i the model expectation at each response χ_{ni} and the subtracting this expectation from the $\chi_{ni} = 0$ or 1 which was actually observed.

The model expectation for χ_{ni} is Π_{ni} where the Rasch model for Π_{ni} is $\Pi_{ni} = \exp(B_n - D_i) / [1 + \exp(B_n - D_i)]$ and B_n and D_i are the parameters which b_n and d_i estimate.

The standardized residual from expectation is

$$z_{ni} = (\chi_{ni} - \Pi_{ni}) / [\Pi_{ni}(1 - \Pi_{ni})]^{1/2}. \quad 16.19$$

When the data fit the model this standardized residual is distributed with mean zero and variance one. Although the expected sampling distribution of D_i is not normal, we have found that values below -2 and above +2 are useful as indicators of noteworthy misfit.

We estimate Π_{ni} from

$$P_{ni} = \exp(b_n - d_i) / [1 + \exp(b_n - d_i)] \quad 16.20$$

where b_n and d_i are the estimates of B_n and D_i , and then use the sampling distributions for $z_{ni} \approx (\chi_{ni} - P_{ni}) / [P_{ni}(1 - P_{ni})]^{1/2}$ of $z_{ni} \sim N(0,1)$ and $z_{ni}^2 \sim \chi_1^2$ as guidelines for evaluating the extent to which any particular set of data is sufficiently coherent to construct useful measurement.

PERSON FIT

To measure the validity of a person's performance, calculate the sum of squared residuals z_{ni}^2 for that person. When the person's behavior is useful for measurement because their response pattern fits the measurement model, then their sum of square standard residuals will approximate a chi-square statistic

$$\sum_i^L z_{ni}^2 = C_n^2 \sim X_{fn}^2 \quad 16.21$$

with degrees of freedom

$$f_n = (L-1)(N-1) / N \quad 16.22$$

and a mean square statistic

$$V_n = C_n^2 / f_n \sim F_{fn, \infty} \quad 16.23$$

with degrees of freedom f_n and ∞ .

ITEM FIT

To measure the validity of an item's usage calculate the sum of squared residuals z_{ni}^2 for that item. When the item is useful for measurement because the pattern of its responses fits the measurement model, then its sum of squared standard residuals will approximate a chi-square statistic

$$\sum_n^N z_{ni}^2 = C_i^2 \sim X_{fi}^2 \quad 16.24$$

with degrees of freedom

$$f_i = (N-1)(L-1) / L \quad 16.25$$

and a mean square statistic

$$U_i = C_i^2 / f_i \sim F_{fi, \infty} \quad 16.26$$

with degrees of freedom f_i and ∞ .

UNDERSTANDING HOW THE RASCH MODEL WORKS

Now we will reexamine the details of PROX to discover what is accomplished in its application. The PROX formula enables a simple and intuitive approach to understanding how Rasch item calibration and person measurement work.

THE RESPONSE MATRIX

Consider the response matrix:

		Items			Person Scores
		1	n	L	
Persons	N	Response $x_{ni}=1$ for correct $x_{ni}=0$ for incorrect			$\sum_i^L x_{ni} = r_n$
	n				
	N				
Item Scores		$\sum_n^N x_{ni} = s_i$			

When there is no missing data* so that this response matrix is complete, then every row total gives a person score for the same set of items and every column total gives an item score for the same set of persons. These item scores are reported as "P-values" in traditional item analysis. "P-values" are calculated by dividing the item score s_i by the number of persons N so that the P-value for item i becomes $P_i = S_i / N$, the proportion of correct responses to item i .

If the person raw scores r_n and item P-values $P_i = S_i / N$ were linear objective quantifiers of person ability and item difficulty, our work would seem done at this point. Indeed, person raw scores and item P-values are as far as traditional person measurement and item analysis go. As a consequence, most of the research results in educational measurement have been limited to what little can be done when raw counts of right answers are mistaken for measures.

As measures, however, raw counts have two serious drawbacks.

* In the application of PROX (or any other Rasch model procedure), the data need not be complete. Missing data can be accommodated without trouble. This happens because the measurement structure specified by the model needs only enough data to identify a finite estimate for each person and each item. As a result none of the estimation procedures need complete data to obtain good estimates.

RAW SCORES ARE NON-LINEAR

Raw counts are bounded in range between none right and all right. Because of this they cannot represent abilities or difficulties on a linear (interval) scale. But, if they are not linear, then the results of the arithmetic used in statistical analysis become misleading. In order to enable the substantial benefits of statistical analysis, we must transform the non-linear scores into linear measures.

RAW SCORES ARE TEST AND SAMPLE DEPENDENT

Each observed response is the result of a person of some ability attempting an item of some difficulty. Because of this, the magnitudes of item scores s_i and P-values P_i , which are summed over persons, depend on the particular abilities of this particular sample of persons and so are sample dependent.

The magnitudes of person scores r_n , which are summed over items, depend, in turn, on the particular difficulties of this particular set of items and so are test dependent.

In order to make general use of the information about person measures and item difficulties which is contained in the test item response data, we must liberate the numerical representations of person measures and item difficulties from the local effects they have on one another. We must construct objective person measures which are test-free and objective item calibrations which are sample-free.

CONVERTING NON-LINEAR, LOCALLY DEPENDENT ITEM AND PERSON SCORES TO LINEAR, INDEPENDENT ITEM CALIBRATIONS AND PERSON MEASURES

The way the Rasch model linearizes raw scores and frees them of sample and test dependency can be seen in the following two PROX formulae:

For Items:

$$\log \left[\frac{(1 - P_i)}{P_i} \right] * \left[1 + \frac{\sigma^2}{2.7} \right]^{1/2} + M \Rightarrow d$$

[log odds
linearize item
P-values]
[scales out
the sample
variance σ^2]
[adjusts out
the sample $\Rightarrow$
mean H]
[Test-freed
item
calibration]

For Persons:

$$\log \left[\frac{r_n}{(1 - r_n)} \right] * \left[1 + \frac{w^2}{2.7} \right] + H \Rightarrow b_n$$

[log odds
linearize person
raw scores]
[scales out
the test item
variance w^2]
[adjusts out
the test item $\Rightarrow$
mean H]
[Test-freed
person
measurement]

For more detail see the example of PROX item calibration and person measure in Chapter 2 of Best Test Design (Wright & Stone, 1979, pp. 28-45).

MEASUREMENT ESSENTIALS

2nd Edition

BENJAMIN WRIGHT

MARK STONE

Copyright © 1999 by Benjamin D. Wright and Mark H. Stone
All rights reserved.

WIDE RANGE, INC.
Wilmington, Delaware