15. ESTIMATING ITEM CALIBRATIONS AND PERSON MEASURES

INTRODUCTION

In this chapter we will work through a mathematical approach to the estimation of Rasch model item and person parameters (Rasch, 1960). This approach is especially suited to computer implementation and most of the computer programs in use employ versions of the algorithms to be described. The procedure is called UCON, for unconditional maximum likelihood estimation (MLE) (Wright & Panchapakasan, 1969; Wright & Douglas, 1977a; Wright & Stone, 1979; Wright, 1980). The term "unconditional" is used because there is another fully conditional maximum likelihood estimation (FCON) which uses conditional probabilities to estimate item difficulties directly without involving any simultaneous estimation of person abilities (Wright, 1968, 1980; Wright & Douglas, 1977b). FCON has desirable theoretical properties, but it is difficult to implement when there are more than a few items. UCON, on the other hand, approximates the results of FCON closely—and UCON seldom has any trouble giving useful results.

Although calibration of item difficulties is the first stage in the implementation of the model, and, in principle, precedes the measurement of persons, it is convenient to estimate item difficulties and person abilities simultaneously. The analysis of fit is expedited by the computation of expected responses of persons to items so that these expected responses can be compared with the observed responses. These expected responses can be determined most easily when we have simultaneous estimates of item difficulties and person abilities.

The estimation of statistical model parameters is the fundamental step of applied statistics. When we view calibration as a problem in statistical estimation, the question arises as to which estimation procedure to use. There are many estimation procedures: least squares, mean value, minimum chi-square, maximum likelihood. The last procedure, MLE, developed by Ronald Fisher in the 1920's, has a number of useful properties. The Rasch model lends itself to MLE and the useful properties of MLE translate into substantive fundamentals of measurement.

RASCH MLE PROCEDURES

Once a statistical model is specified, an equation for the probability of occurrence of any observation follows. From this equation, the joint probability of any data set may also be specified and this equation used to answer the question: What is the probability that this particular set of data occurred when this set of items was given to this group of persons? This joint probability is known as the likelihood of the data. It is a function of the observed data and also of the initially unknown but soon to be estimated parameters of the model (the item difficulties and person abilities). The MLE principle is to select for the estimates of the parameters that particular set of values which makes the likelihood of the data in hand as large as possible - a maximum.

The likelihood of the data is viewed as a function of known data and unknown parameters. The parameters become the variables. Calculus is employed to find the particular values of these unknown parameters that make the likelihood of these data a maximum. This is done by taking the derivative of the likelihood with respect to each unknown variable and setting this derivative equal to zero. This produces equations which may be solved for the unknown values, which, when obtained, make the likelihood of these data as large as it can get.

To review, we:

1) derive an expression for the likelihood of the data,

2) differentiate this expression with respect to each of the unknown parameters,

3) set each result equal to zero and

4) solve the resulting set of equations for the ML item difficulty and person ability estimates.

When the Rasch model is applied to test data there are a large number of unknown parameters to be estimated, many more than the one or two involved in the usual maximization problem. Nevertheless, the principles are the same and when the procedure is applied step-by-step to one item and then one person at a time no complications arise.

Usually when we solve equations for an unknown value in algebra, arithmetic operations like addition and division are sufficient to obtain an explicit solution. The equation $5X+6=20$, for example, requires one subtraction and one division to reach the exact solution of $X = 2.8$. Since this kind of equation can be solved by a finite number of simple arithmetical steps, it is called explicit.

In contrast, an equation like $X + 2 * \sin X = .73$ does not lend itself to simple arithmetic. To solve this "implicit" equation we must resort to another method. A good way to solve this kind of implicit equation was invented by Isaac Newton in the 1680's.

Newton's method:

1) a reasonable guess is provided for the unknown value of X,

2) the "closeness" of this guess to the best solution is determined by noting how much remains when this value for X is substituted in the equation,

3) the difference between the initial value for X and the remainder is then used to determine a next "better" value for X,

4) this process for improving the estimate of X continues until the remainder gets small. How small is left to the discretion of the person solving the equation.

Each step in this process is called an iteration. The iterative process will converge to a solution for a large class of implicit equations, among which are equations incorporating the exponential function $\exp(X)$. All that are needed to implement Newton's method are the derivatives of the equations to be solved and good initial guesses. For the Rasch model equations, there are very sensible initial guesses for the unknown item difficulties and person abilities.

If we let $f(x) = 0$ be the equation to be solved for the unknown X, and, if $f'(X)$ is its derivative with respect to X and, if X_0 is the initial guess for the value of X, then Newton's method specifies the next better value for X as

$$X_1 = X_0 - \frac{f(X_0)}{f'(X_0)} \qquad \textit{15.1}$$

where $f(x_o)$ and $f'(x_o)$ are values of these functions when we substitute the initial value x_o for X and x_t is the new, improved value for X at the end of the first iteration.

We can write a general expression for this relation which shows the value of x_t at the end of t iterations in terms of what it was on the previous iteration:

$$X_t = X_{t-1} - \frac{f(X_{t-1})}{f'(X_{t-1})} \qquad \textit{15.2}$$

Since we may continue iterating until our result is as accurate as we wish, when should we stop when estimating parameters for a Rasch model? Experience has shown that when reporting values for item difficulties and person abilities we never need accuracy greater than two decimal places. Enough accuracy is obtained when we settle for an x_t which makes the absolute difference between that x_t and its previous value x_{t-1} in the vicinity of 0.005, that is, "correct" to the second decimal place.

MAXIMUM LIKELIHOOD ESTIMATION

The Rasch probability of any observation x_{ni} for person n on item i is

$$P\left(X_{ni} \middle| B_n,\ D_i\right) = P_{ni} = [\exp\ X_{ni}(B_n - D_i)] \,/\, [1 + \exp(B_n - D_i) \qquad \textit{15.3}$$

where x_{ni} is the observed data, and may be either 0 or 1,

B_n is the unknown person ability measure and

D_i is the unknown item difficulty calibration.

For a test of L items given to N persons for whom it is reasonable to think of the persons and items as functioning independently i.e. as specified by *Equation 15.3*, the joint probability (the likelihood) of all the data is found by multiplying together all N by L probabilities of the type in *Equation 15.3.*

The expression $(A**m)*(A**n)*(A**q)$ may be written with the single base A and an exponent which is the sum of the three exponents, $A**(m+n+q)$. When this notation is applied to the N x L exponents of the likelihood function, we have

$$\wedge = \prod_i^L \prod_n^N P_{ni}$$

$$= \prod_{i=1}^L \prod_{n=1}^N \left[\frac{\exp[X_{ni}(B_n - D_i)]}{1 + \exp(B_n - D_i)} \right]$$

$$\left[\frac{\exp\left[\sum_i^L \sum_n^N\right](X_{ni}B_n - X_{ni}D_i)}{\prod_i^L \prod_n^N [1 + \exp(B_n - D_i)]} \right] \qquad 15.4$$

where is the likelihood of the data, $\prod_i^L \prod_n^N$ is the continued product over n and i of all $N * L$ probabilities P_{ni} and $\sum_i^L \sum_n^N$ is the continued sum over n and i of all $N * L$ exponents $(X_{ni}B_n - X_{ni}D_i)$.

The double summation in the numerator can be distributed over the two terms with the result

$$\wedge = \left[\frac{\exp\left[\sum_n^N B_n \sum_i^L\right] X_{ni} - \sum_i^L D_i \sum_n^N X_{ni}}{\prod_i^L \prod_n^N [1 + \exp(B_n - D_i)]} \right]$$

$$= \frac{\exp\left[\sum_n^N B_n R_n - \sum_i^L D_i S_i\right]}{\prod_i^L \prod_n^N [1 + \exp(B_n - D_i)]} \qquad 15.5$$

where $\sum_i^L X_{ni} = R_n$ is the right answer count or the raw test score for person n,

and $\sum_n^N X_{ni} = S_i$ is the right answer count or raw sample score for item i.

With the likelihood in this form we see that the statistics required are not the separate person-to-item responses but only their accumulations into the person scores R_n and the item scores S_i. Further, the R_n's and S_i's are separated from each other. Each set multiplies its own parameters B_n's and D_i's in turn. This separation is the defining characteristic of a Fisher "sufficient" statistic (Fisher, 1958) and also the algebraic requirement for Rasch objectivity.

Although R_n and S_i are sufficient to estimate B_n and D_i these scores themselves are not satisfactory as measures. Person score is not free from the particular item difficulties encountered in the test. Nor is item score S_i free from the ability distribution of the persons who happen to be taking the item. Independence from these local factors requires adjusting the observed R_n and S_i for the item difficulty and person ability distributions they depend on. This adjustment is necessary to produce the test-free person measures and sample-free item calibrations we desire.

In order to obtain the maximum of this likelihood with respect to possible values of the unknown parameters, the likelihood needs to be differentiated with respect to the B's and D's in turn. This task is easier when we take the logarithm of the likelihood. We can do that because the values which make the logarithm of a function a maximum also make that function a maximum.

Since $\log(\exp X) = X$, the numerator of the log likelihood becomes simple. The denominator turns into a subtraction and the double product becomes a double sum of $\log\left[1+\exp(B_n - D_i)\right]$.

Thus $K = \log \wedge = \sum_n^N B_n R_n - \sum_i^L D_i S_i - \sum_i^L \sum_n^N \log\left[1+\exp(B_n - D_i)\right]$ is the log-likelihood. *15.6*

Since the derivative of the exponential function, $\exp X$, reproduces itself and the derivative of the logarithmic function, $\log Y$, is $1/Y$, the differentials required to produce solutions for $\partial K / \partial B$ and $\partial K / \partial D$ are

$$\partial (D_i S_i) / \partial D_i = S$$

$$\partial (B_n R_n) / \partial B_n = R_{ni}$$

$$\frac{\partial \log\left[1+\exp(B_n - D_i)\right]}{\partial B_n} = \frac{\exp(B_n - D_i)}{1+\exp(B_n - D_i)} = P_{ni}$$

$$\frac{\partial \log\left[1+\exp(B_n - D_i)\right]}{\partial D_i} = -\frac{\exp(B_n - D_i)}{1+\exp(B_n - D_i)} = -P_{ni} \qquad 15.7$$

By differentiating the log likelihood K with respect to each D_i and then, separately, with respect to each B_n and equating each of these derivatives to zero to locate maxima, we obtain the two sets of equations.

$$\frac{\partial K}{\partial D_i} = -S_i + \sum_n^N \frac{\exp(B_n - D_i)}{1+\exp(B_n - D_i)}$$

$$= -S_i + \sum_n^N P_{ni} = 0 \text{ for each } i = 1, L$$

$$\frac{\partial K}{\partial B_n} = +R_n - \sum_i^L \frac{\exp(B_n - D_i)}{1+\exp(B_n - D_i)}$$

$$= +R_n - \sum_i^L P_{ni} = 0 \text{ for each } n = 1, N \qquad 15.8$$

Each of the first L equations contains N unknown B_n's and one unknown D_i. Each of the second N equations contains L unknown D_i's and one unknown B_n.

Newton's method uses the derivative of the equation to be solved, therefore we need to take the derivatives of the above implicit equations with respect to D_i and B_n once again in order to solve them by Newton's method. These derivatives are the second derivatives of the likelihood.

Since $$P_{ni} = \frac{\exp(B_n - D_i)}{1 + \exp(B_n - D_i)} = \frac{1}{1 + \exp(D_i - B_n)} \qquad 15.9$$

the differentials needed to find the second derivatives of K with respect to B_n and D_i are

$$\frac{\partial K}{\partial D_i} = -S_i + \sum_n^N P_{ni}$$

$$\frac{\partial^2 K}{\partial D_i^2} = -\sum_n^N P_{ni}(1 - P_{ni}) = -\sum_n^N Q_{ni} \text{ for } i = 1, L$$

$$\frac{\partial K}{\partial B_n} = R_n - \sum_i^L P_{ni}$$

$$\frac{\partial^2 K}{\partial B_n^2} = -\sum_i^L P_{ni}(1 - P_{ni}) = -\sum_i^L Q_{ni}$$

$$\text{for } n = 1, N \text{ where } Q_{ni} = P_{ni}(1 - P_{ni}) \qquad 15.10$$

These second derivatives are the product of P_{ni} and its complement $(1-P_{ni})$ combined in $Q_{ni} = P_{ni}(1 - P_{ni})$ where P_{ni} is the probability that person n gets item i correct.

Since $Q_{ni} \geq 0$, these second derivatives are always negative. This tells us that the solutions to *Equation 15.8* must be maxima.

Before we apply Newton's method to solve these equations, three uncertainties need to be resolved.

1. What shall we use for initial values of the estimates? Although Newton's method is usually robust with respect to the choice of an initial estimate (meaning we will get to the same final estimate no matter where we start), we will get convergence most rapidly if we use initial estimates which are not far from the final estimates.

   We can do this for items, by approximating the abilities of all persons at zero.

Then the MLE's of *Equation 15.8* have the explicit solution:

$$\sum_{n}^{N} P_{ni} = \sum_{n}^{N} \frac{\exp(-D_i)}{1+\exp(-D_i)} = \frac{N\exp(-D_i)}{1+\exp(-D_i)}$$

so
$$-S_i + \sum_{n}^{N} P_{ni} = -S_i + \frac{N\exp(-D_i)}{1+\exp(-D_i)} = 0$$

$$\exp(-D_i) = S_i / (N - S_i)$$

and
$$D_i = -\log\left[\frac{S_i}{N-S_i}\right] = +\log\left[\frac{N-S_i}{S_i}\right] \text{ for } i = 1, L \qquad 15.11$$

This initial estimate is a simple logarithmic transformation (the logit) of the item scores.

By approximating the difficulties of all items at zero in the equations for the B's in *Equation 15.8*, we find a similar explicit solution for B_n as a simple logarithmic transformation of the raw scores for person n

$$B_n = \log\left[\frac{R_n}{L-R_n}\right] \text{ for } n = 1, N \qquad 15.12$$

2. Although it may appear that the equations in 15.8 have N unknowns $B_1, B_2, \ldots, B_N$ only the statistics $R_1, R_2, \ldots, R_{L-1}$ are available to estimate them. When data is complete the values of B_n which can be estimated from a test of L items may therefore be indexed by R rather than by n. Indexing persons by their raw scores highlights the fact that a raw score for a person is the sufficient statistic for estimating that person's ability.

In general, there will be more than one person with a given raw score. Since as far as ability estimation is concerned, we are unable to distinguish among persons who took the same items and earned the same raw score. We may group persons who took the same items according to their raw score. If we let N_r be the number of persons who scored R on the test, we may rewrite *Equations 15.8* and *15.10* as

$$-S_i + \sum_{R=1}^{L-1} N_R P_{Ri} = 0 \qquad 15.13$$

and
$$\frac{\partial}{\partial D_i}\left(\frac{\partial K}{\partial D_i}\right) = -\sum_{R}^{L-1} N_R P_{Ri}(1 - P_{Ri})$$

$$= -\sum_{R}^{L-1} N_R Q_{Ri} \qquad 15.14$$

and *Equation 15.12* as $$B_R = \log\left[\frac{R}{L-R}\right] \qquad 15.14$$

$$R=1,\ L\text{-}1$$

where $$P_{Ri} = \exp(B_R - D_i)/\left[1+\exp(B_R - D_i)\right]$$

and $$Q_{Ri} = P_{Ri}(1-P_{Ri})$$

3. Were we to apply Newton's method to these equations as they stand, we would find that the iteration process would not converge. This is because our set of equations contains one too many unknowns to be uniquely estimated.

The Rasch model specifies the probability of a response by a person to an item as a function of the difference between their locations on a variable.

The probability that a person with ability B_n gets an item with difficulty D_i correct, is exactly the same as the probability for a person with ability (B_n+3), say, responding to an item with difficulty (D_i+3), because $(B_n+3)-(D_i+3)=B_n-D_i$. Since our choice of 3 was arbitrary, we see that an infinite set of B's and D's will satisfy our equations providing only that they maintain their differences $(B_n - D_i)$.

This problem of too many unknowns can be overcome by placing one restriction on the set of B_n's and D_i's. The particulars of this restriction are not important algebraically. We could set any person, say B_1, equal to a constant or any item, say D_3, equal to some other constant. Any constant will do. We have found it convenient for calibration to use the restriction that the sum of our set of estimated item difficulties $\sum_i^L D_i \equiv 0$ be zero. This centering on the test has the effect of reducing our unknowns from $(L-1)+L=2L-1$ to $(L-1)+(L-1)=2L-2$.

In order to maintain the possibility of convergence, we must implement this restriction each time we derive an improved set of (D_i) values. Centering is accomplished by finding the mean of the current estimates of the D_i's and subtracting this mean from each D_i. This is done at each iteration.

Thus the initial centered set of D_i's are

$$D_i = \log\left[\frac{N-S_i}{S_i}\right] - \sum_i^L\left\{\log\left[\frac{N-S_i}{S_i}\right]\right\}/L \qquad 15.16$$

SOLVING THE MAXIMUM LIKELIHOOD EQUATIONS

Here is a systematic procedure for solving these equations and hence obtaining estimates of item difficulty and person ability (Once all perfect and zero scores have been removed from the data matrix).

1. Determine the initial item estimates from *Equation 15.16.* Items are centered.

2. Determine the initial person estimates from *Equation 15.15.* Persons do not need to be centered. In fact, they must not be.

3. Using all person estimates and the current estimate for each item i, apply Newton's method to *Equation 15.13* until differences between successive estimates of each D_i that is, $(D_i' - D_i)$ are less than, say, .005 logits. The process is,

$$D_i' = D_i - \frac{S_i - \sum_R^{L-1} N_R P_{Ri}}{\sum_R^{L-1} N_R Q_{Ri}} \qquad 15.17$$

   in which D_i is the current estimate and D_i is the next improved estimate. The successive differences are $(D_i' - D_i)$.

4. Repeat step 3 for all items, $i = 1, L$. When we have finished, we have a new and better set of D_i estimates.

5. Center these new D estimates.

$$D. = \sum_i^L D_i \,/\, L \qquad\qquad D_i' = D_i - D. \qquad 15.18$$

6. Using these new centered D_i estimates and the person estimate for a score of $r = 1$ (that is, B_1), apply Newton's method to *Equation 15.10* expressed in terms of r instead of n.

$$r - \sum_i^L P_{ri} = 0 \qquad 15.19$$

   until differences between successive estimates of B_r that is $(B_r' - B_r)$ are less than, say, .005 logits.

   The process is,

$$B_r' = B_r + \frac{r - \sum_i^L P_{ri}}{\sum_i^L Q_{ri}} \qquad 15.20$$

in which B_r is the current estimate and B_r' is the improved estimate. The successive differences are $(B_r' - B_r)$.

7. Repeat step 6 for the second, then third, etc., raw score. When we have reached $r = L - 1$ we have a new and better set of B_r estimates. Do *not* center these B_r.

8. At this stage we have reached the end of the first major "loop". This loop comprised L minor loops on the items and L-1 minor loops on the person scores.

At the end of each major loop we determine whether the likelihood has been sufficiently maximized by reviewing our convergence criterion for all $2L$-1 estimates. Since it is unlikely that satisfactory convergence will have been achieved in one major loop, we proceed to additional major loops.

9. Using the latest person estimates and the current value for D_1, apply Newton's method as in Step 3 until convergence. Repeat for all items.

10. Center the latest set of D_i's.

11. Using these latest centered D_i and the current value for B_1, apply Newton's method, Step 6, until convergence. Repeat for all raw scores from 2 to L-1.

12. Determine whether a satisfactory overall convergence has been obtained at the end of this second major loop and so on.

This estimation procedure usually converges to a criterion of .005 logits in 5 or 6 major loops. There are rare circumstances in which an MLE cannot be obtained. When there are one or two items or persons separated from the nucleus of the data by many logits, then round-off problems can occur with the procedure outlined above. When this procedure fails it is almost always due to inaccurate editing of the original data or to failure to center items each time a new set of estimates is produced.

Because of the way estimates are calculated in UCON there is a slight bias. This bias can be corrected by shrinking all values of B and D by the factor $(L$-1$)/L$ (Wright, 1988).

STANDARD ERRORS OF ESTIMATES

A key benefit of a good estimation procedure is the simultaneous estimation of standard errors for its estimates. These standard errors specify the modeled degree of precision (reliability) with which the estimates can be obtained.

A familiar example of this is estimating a mean from a random sample of N observations. The sample mean is, in many ways, a "best" estimate of the location parameter of the distribution from which the random sample was drawn. The standard error of a mean is given by $S / N^{1/2}$, where S is an estimate of the dispersion of the distribution calculated from the standard deviation of the N observations. Notice that this standard error, or precision of estimation, is dominated by the size of the sample N; the larger the sample size, the smaller the standard error and so the greater the precision.

With respect to the MLE procedure just described for the Rasch model, Ronald Fisher proved that as long as sample size is reasonably large, the standard error of a ML estimate is well estimated by the inverse negative square root of the second derivative of the likelihood function.

Fisher also proved that replicates of a MLE will, with sufficiently large sample size, have a normal distribution with expected value equal to the parameter itself and with a standard deviation equal to this standard error. We will use this result to set confidence limits on our MLE's.

The second derivative of the likelihood, which served as a scaling factor for Newton's method, now plays an important statistical role. It gives us the standard errors for our estimates.

Thus from *Equations 15.10* and *15.14*

$$SE(D_i) = \left[\frac{-\partial}{\partial D_i} \left(\frac{\partial K}{\partial D_i} \right) \right]^{-1/2} = \left(\sum_r^{L-1} N_r Q_{ri} \right)^{-1/2} \sim 2.5 / N^{1/2} \qquad 15.21$$

$$SE(B_r) = \left[\frac{-\partial}{\partial B_r} \left(\frac{\partial K}{\partial B_r} \right) \right]^{-1/2} = \left(\sum_i^{L} Q_{ri} \right)^{-1/2} \sim 2.5 / L^{1/2} \qquad 15.22$$

These standard errors are determined by substituting into $P_{ri} = \exp(B_r - D_i) / [1 + \exp(B_r - D_i)]$ the converged values of the B_r and D_i estimates and finding $Q_{ri} = P_{ri}(1 - P_{ri})$.

Item calibration and person measurement is now complete. Here is a summary of the results of this MLE.

A set of L items estimates D_i is obtained whose sum has been set to zero so that the measuring system under construction has been centered on the calibrations of these test items. Some values will be negative, indicating relatively easy items and some will be positive, indicating relatively hard items.

Associated with each of the D_i estimates is its estimated standard error. We will see that D_i's with values far from the sample of persons have relatively large standard errors and that standard errors get smaller (and hence items more precisely estimated) as we get closer to items with D_i values near the B_r values of the majority of the persons. This is a consequence of the formula for the standard error in *Equation 15.21*. When D_i is equal to B_r, the value of P_{ri} is 0.5 and so Q_{ri} is 0.25, its maximum possible value. This is where the standard error is nearest to its theoretical minimum of $2 / N^{1/2}$.

Equation 15.22 shows the way the standard error of B_r is a function of the number of items near that B_r. The standard errors of the person abilities depend on how many of the item difficulties are near the location of that person. The more items near the person, the smaller the standard error of the measure. In a test which is well centered on its target group we would expect the standard errors of person abilities to be symmetric around a central B near zero, corresponding to about half the items correct, and that these standard errors would be large at both ends of our variable line and get smaller towards the center.

Since MLE produces values for all scores r, for $r = 1$ to $r = L - 1$, we will have ability estimates and standard errors for all possible scores, even when, in the calibrating sample, there were no persons who actually obtained a particular score.

MEASUREMENT ESSENTIALS

2nd Edition

BENJAMIN WRIGHT

MARK STONE

WIDE RANGE, INC.
Wilmington, Delaware