The Point is Measurement

In spite of most of what has been said up to this point, we did not undertake this project with the intent of building better thermometers. The point is to measure the person. Because of the complete symmetry of the model, everything we have done for items, we can do again for people just by reversing the subscripts. For any two people who took some of the same items, count the number N_{12} that person *2* answered correctly and person *1* missed; also the number N_{21} that person *1* passed and person *2* missed. The relative abilities of the people will parallel expressions 23 and 25:

33. $\frac{B_2}{B_1} = \frac{N_{12}}{N_{21}}$, or its reciprocal.

34. $b_1 - b_2 = \ln(N_{21}) - \ln(N_{12})$ or its negative.

It is not necessary that the people take all the same items, just so there is some overlap for the pair. This can be generalized to any number of people by building the bigger matrices (see Tables III.2 and III.3) and solving the simultaneous equations. But, there are some reasons we don't want to do this.

- It would be awkward (and not pure Raschian) to assign different Scale Scores to people with the same raw score.
- People who missed very few items (or passed very few) give us almost no useful data to work with[1].
- The matrices will have one row and one column for each person tested and quickly become unwieldy for even a modest-sized assessment.

However, we know that the number correct score is the sufficient statistic for ability; we don't need a score for every person, just for every number correct score. The issues can be circumvented or at least mitigated by indexing the people by their raw scores and tabulating the results as though everyone with the same score is the same person. On the other hand, if we know the item difficulties well enough, we do what we have done since Panchapakesan (1969).

Counts to Measures

A more computationally efficient process that has been the workhorse of US-based Rasch analysis (Wright and Panchapakesan, 1969) assumes the difficulties are known (i.e., good enough estimates of the difficulty parameters are available); no additional data are needed. The ability estimate b_r associated with the raw score r is the value that satisfies the basic equation:

35. $$r = \sum_{i=1}^{L} P(x_{ri} = 1) = \sum_{i=1}^{L} \frac{B_r}{B_r + D_i} = \sum_{i=1}^{L} \frac{e^{b_r - d_i}}{1 + e^{b_r - d_i}}.,$$

where r is a raw score from *1* to *L-1*; L is the total number of items; $P(x_{ri})$ is the probability of a correct response on item i for a person with the ability b_r associated with

[1] No approach to estimation has a particularly satisfying answer to the question of what to do with the people with zero or perfect scores. The same issue exists for items but it is easier to ignore. We will return to the topic shortly with a couple of contrived suggestions.

r. Because total raw score is the sufficient statistic for estimating ability, everyone who took the same items and got the same raw score gets the same estimated ability b_r. Hence the probability and estimate can be indexed by the raw score, instead of the person.

Equation 35 simply says the expected total score Σp_{ri} is equal to the observed total score r; if they aren't equal enough, the ability estimate needs adjusting. If the expected score is low, the estimated ability is increased; if the expected score is too high, the estimate is decreased. The d_i are taken to be known. We only need to fiddle with the b_r until the equation is true. That's all there is to computing abilities; the rest is details for doing the fiddling.

The ability estimate is adjusted by your favorite numeric method until equation 35 is satisfied. Wright & Panchapakesan (1969) applied Newton's method to do the iterating:

$$36. \qquad b_r^{t+1} = b_r^t + \frac{r - \sum_{i=1}^{L} p_{ri}}{\sum_{i=1}^{L} p_{ri}(1-p_{ri})} .$$

An effective starting value for this process is:

$$37. \qquad b_r^0 = \ln\left(\frac{r}{L-r}\right) - \bar{d} \quad \text{where } \bar{d} = \frac{\sum_{i=1}^{L} d_i}{L} \text{ is the center of the item difficulties,}$$

which is often zero.

Table III.7 shows the arithmetic for a small test with 10 dichotomous items. It is typical in this situation for the process to stabilize in two or three iterations. This process for estimating ability can be derived with standard maximum likelihood methods: define a likelihood function for the data; take the first derivative; set equal to zero; solve and but first check that you've got a maximum not a minimum. That's all the detail I'm going to give, but I will note that the symmetry of the model means you can turn the notation around and do exactly the same thing for items.

Table III.7: Calculations of Logit Abilities for a Test with 10 Dichotomous Items

Item Logit	Raw Score	Initial	Round One	Round Two	Std Error
	0			-3.494	1.74
0.637	1	-2.197	-2.339	-2.347	1.071
-0.941	2	-1.386	-1.496	-1.499	0.814
-0.266	3	-0.847	-0.922	-0.923	0.716
0.382	4	-0.405	-0.444	-0.444	0.674
-0.455	5	0	-0.001	-0.001	0.661
0.086	6	0.405	0.441	0.442	0.674
-0.881	7	0.847	0.920	0.921	0.717
0.000	8	1.386	1.496	1.499	0.815
0.297	9	2.197	2.341	2.349	1.073
1.141	10			3.500	1.74

Item Logit	Raw Score	Initial	Round One	Round Two	Std Error
	Sum of p				
	1	1.138	1.007	1	
	2	2.175	2.005	2	
	3	3.149	3.002	3	
	4	4.085	4		
	5	5.003	5		
	6	5.92	6		
	7	6.854	6.998	7	
	8	7.826	7.995	8	
	9	8.86	8.993	9	
	Sum of p(1-p)				
	1	0.972	0.876	0.871	
	2	1.601	1.513	1.510	
	3	1.997	1.949	1.948	
	4	2.217	2.204	2.204	
	5	2.287	2.287	2.287	
	6	2.215	2.202	2.202	
	7	1.993	1.945	1.945	
	8	1.596	1.509	1.506	
	9	0.971	0.874	0.869	

All Right or All Wrong

Equation 37 for the starting value makes it obvious, but it also follows more subtly and more profoundly from the estimation equation 35, that perfect scores, both *r=0* and *r=L*, are problems. There is no ability low enough to ever satisfy equation 35 when *r* is *0*, nor high enough when *r* is *L*. In the real world, it is generally necessary to manufacture something to report for examinees with these scores, although it would be much preferred to avoid giving tests so far off target. One tactic is to solve the equations for non-integer scores arbitrarily close to the perfect scores, say, within 0.25. Whether the target should be off by 0.25, or 0.1, or 0.33, or some other value is completely arbitrary; the smaller the value, the more extreme the solutions will be. It is more a policy decision than psychometric issue about how much punishment or reward should be attached to those scores.

Another strategy, with slightly more psychometric underpinning and avoids the arbitrary choice of target, produces almost the same results by assigning to a score of zero the *logit ability for a raw score of one minus its squared standard error of measurement*:

38. $b_0 = b_1 - s_1^2$.

Analogously for a perfect score of *L*, the logit ability estimate is the estimate for a score of *L-1* plus its squared standard error. The simple rationale for this tactic is that the

difference between logit ability estimates for any adjacent scores is very nearly equal to the squared standard error of measurement. The more erudite explanation is that, because the squared standard error is the inverse of the denominator of expression 36[2], this tactic is equivalent to using expression 36 to estimate the ability for zero (or *L*) using the starting value b_1 (or b_{L-1}) and stopping after the first iteration.

This is the method used in Table III.7, although three decimals implies more precision than I feel about this step.

39. $b_0 = -2.347 - (1.071)^2 = -3.494.$

40. $b_{10} = 2.349 + (1.073)^2 = 3.500.$

For tests with dichotomous items, the standard errors for 1 and L-1 will typically be a little more than one. Squaring that gives about *1.15* or *1.2.* Using either of these values in place of s_1^2 or s_{L-1}^2 gives almost the same result as either of the other methods. We're making the numbers up anyway but that's too simple to be given serious scholarly consideration. That's the trouble with Rasch.

Standard Errors of Measurement

A statistician is a person with a bag of standard errors and who can produce the appropriate one for any situation. Theodore Bancroft

The Pair algorithm has been criticized for the lack of an asymptotic standard error estimator and who wouldn't want one of those. That doesn't mean that we don't have a suggestion. A reasonable possibility for the standard error for each element of the ***R*** matrix is:

41. $$s_{ij}^2 = \frac{1}{4}\left[\frac{1}{n_{ij}} + \frac{1}{n_{ji}}\right] = \frac{n_{ij} + n_{ji}}{4 n_{ij} n_{ji}}.$$

It would be aggregated to the row average as:

42. $$s_i^2 = \frac{\sum_{j \neq i} s_{ij}^2}{L^*}$$

where L^* is the number of defined elements in row *i* of *R* and the factor *1/4* arises because of a lack of independence in ***R***. Because every item response can influence several item pairs, the counts are not independent and hence best case values. But at least the estimates won't be walking around naked and unchaperoned.

For the so-called marginal maximum likelihood estimation process we used to estimate abilities, we do have an estimate of the asymptotic standard error for the logit ability at each raw score. For the simple case of dichotomous items, the standard error for the ability estimate at a raw score *r* is:

[2] The numerator will be plus or minus one because we are using the ability estimate for the score that is one off from where we want to be.

43. $s_r = 1 \Big/ \sqrt{\sum_{i=1}^{L} p_{ri}(1-p_{ri})}$.

These apply to the estimates produced by expression 35 and are sometimes referred to as *conditional* standard errors to distinguish them from *The Standard Error* of true score theory but at this point I prefer to think of them as functions of *r* rather than conditioned on *r* and forget all about the thing they have supplanted.

Standard Error of Measurement; Not Standard Error of Score

The standard error function $s_r = 1 \Big/ \sqrt{\sum_{i=1}^{L} p_{ri}(1-p_{ri})}$ defines a bowl-shaped curve, meaning we are more confident of our measurements near the center of the test than at the extremes. Some find this upsetting because it seems to run counter to what they were taught in their formative years. *True score* theory tells us to have more confidence in scores at the extremes than in the center of the score range, i.e., a dome-shaped function of the form $\sqrt{\sum_{i=1}^{L} p_i(1-p_i)}$.

There really is no inconsistency; we are talking about two different standard errors. One is the standard error for a measure and the other is the standard error for a score. If we give a test that is much too easy, we have a very good idea what a person's score will be: perfect or very near to it and hence a small standard error for the score. But a perfect score is consistent with a very large range of abilities, from here to infinity; hence, a huge standard error for the measure.

At the extremes of the score ranges, we know what the "true" scores must be but have very little idea what the abilities are. Conversely, near the center of a test, we have the least confidence in the number correct score and the most confidence in the logit measure. In the modern world, $\sum_{i=1}^{L} p_{ri}(1-p_{ri})$ should be thought of as the *information function*, which is maximized at the center of the test. No one seriously thought giving an off-target test was a good idea.