

# HOW DO RASCH AND 3P DIFFER?

## An Example from the National Adult Literacy Study

John M. Linacre & Benjamin D. Wright  
MESA Laboratory, 5835 Kimbark, Chicago 60637  
June 13, 1995

When you encounter a complicated data set, do you ever wonder whether the ETS Birnbaum 3 item parameter model [3PL] would do better than the far simpler Rasch model? The 1992 National Adult Literacy Survey [NALS] rib-spiralled 24,944 adults through a survey of 173 literacy items to produce an 80% empty 173 x 24,944 matrix of dichotomies.

ETS reports a strenuous 3PL analysis of these data: difficulties and discriminations for all 173 items and "guessing" lower asymptotes for 13 MCQ items. ETS "measures" each of the 24,944 respondents with 15 "plausible values", randomly selected from posterior ability distributions. The means of these 15 values are analogous to measures. Their S.D.s are analogous to S.E.s.

MESA used BIGSTEPS for 90 minutes on a laptop to analyze the same data. Each respondent got 1 Rasch literacy measure with standard error and fit evaluation. Each item got 1 difficulty calibration with its own standard error and fit evaluation. Comparison of Rasch results with the more complex and expensive 3PL analysis provides an empirical test of the degree to which 3PL might be more informative than Rasch.

When the 15 3PL "plausible values" and the 1 Rasch measure are correlated over the 24,944 adults and the  $(16 \times 15)/2 = 120$  raw correlations disattenuated for measurement error, no correlation is less than .92. 40 of the 120 exceed .99.

Principal component analysis of the 16 x 16 raw correlations finds a 1st factor absorbing 88% of the total variance and a 2nd factor of less than 3%. The 15 "plausible values" contribute absolutely nothing but unnecessary profusion to the measurement of adult literacy.

3PL, in notation and practice, however, is much less interested in what they refer to as the "incidental" person parameters than in its 3 item parameters. Are there any useful differences between the Rasch and 3PL item results?

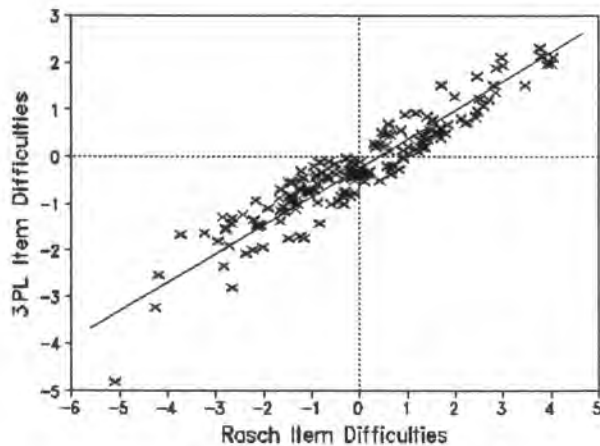
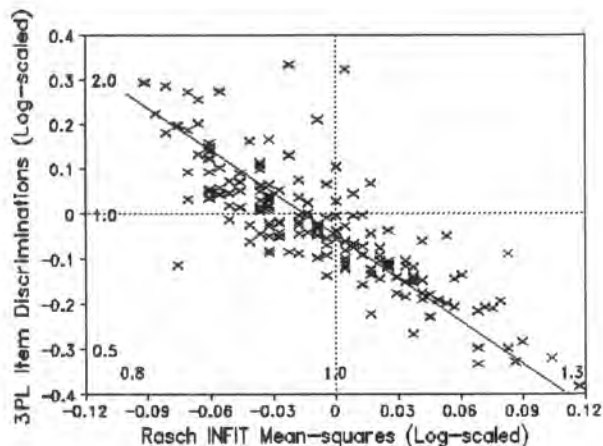


Figure 1 shows Rasch and 3PL item difficulties correlating .94. Since 3PL difficulties are inevitably perturbed by the attempt to estimate item discriminations from the same data, .94 is high.

Rasch specifies that all items have the same discrimination. It is this specification which obtains the parameter separation necessary for conjoint additivity, sufficient statistics and the uncrossed ICCs necessary to build a coherent construct definition.

The slopes of empirical ICCs always differ. The inferential question is how to handle this differing? Shall we mistake it for an enduring item characteristic to be carried forward in parameter form. Or shall we recognize it as a local description of how this particular sample took to this item?

Conventional raw score item analysis evaluates variation in ICC slope as a sample dependent point biserial fit statistic. So does Rasch. But, instead of target dependent point biserials, Rasch uses mean square ratio fit statistics.



To find the 3PL discriminations in a Rasch analysis one need only examine Rasch item INFIT statistics. Figure 2 shows that 3PL item discrimination (log-scaled) and Rasch item INFIT mean-square (log-scaled) correlate -.82. They contain exactly the same information:

$$\log(3PL \text{ discrimination}) \approx -3.3 * \log(INFIT \text{ Mnsq}).$$

The only difference is that 3PL attempts to use this information as a reproducible characteristic of the item, an inference, while Rasch recognizes its unstable sample dependence as no more than a passing description of this occasion.

Nearly all NALS item INFIT mean-squares meet the usual fit criteria (0.8 - 1.2). Most of the ETS differences in discrimination can be attributed to random variation. Parameterizing discrimination does not clarify item function. Instead, it introduces an unreplicable and unnecessary stochastic component.

Some researchers fear the possibility of lucky guessing favoring low performers. ETS attempted to counteract item guessability in the NALS data by estimating a third, lower asymptote, item parameter for 13 NALS MCQ items. Instead of attributing respondent guessing to items, Rasch routinely reviews person outfit to detect lucky guessing when and by whom it occurs.

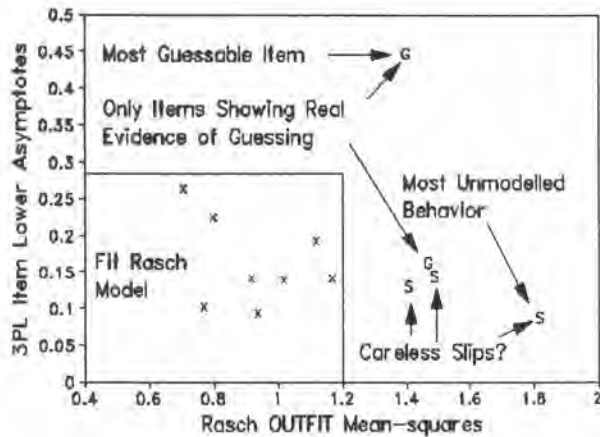


Figure 3 shows that when 3PL lower asymptotes are plotted against Rasch item OUTFIT mean-squares almost no NALS guessing occurred. Only 2 of the 13 MCQ items, "G" in Figure 3, show evidence of guessing. In contrast, 3 other MCQ items show evidence of carelessness (unexpected failure by high ability respondents), "S" in plot.

In fact the NALS data contain more unexpected failures than unexpected successes. Parameterizing lower asymptotes for NALS data is contraindicated.

The large, complex NALS data, containing a wide variety of dichotomous item types and administered to a large and diverse sample of respondents, is just the data expected to manifest all the features that would make the supposed "superiority" of 3PL clear. This parallel analysis shows, however, that 3PL has no benefits over Rasch and some detriments. 3PL ability estimates and item difficulties are identical to Rasch measures. 3PL item discrimination provides no more information than Rasch INFIT statistics. But parameterizing item discrimination complicates estimation and obscures the interpretation of item hierarchy necessary for stable construct definition.

Estimating lower asymptotes is also detrimental. Adding this third item parameter penalizes all respondents, particularly low performers who "know" the right answer. Should probable guessing be detected by person (or item) misfit, a simple strategy is to remove the putative lucky guesses from the data, treating those few items as not administered to those few people. Then, only those who guessed are penalized and only by the small amount by which their lucky guesses boosted their performances.