RASCH MEASUREMENT

Transactions of the Rasch Measurement SIG
American Educational Research Association

Vol. 27 No. 2 Autumn 2013 ISSN 1051-0796

A Critique of 3-PL IRT Estimation

Ben Wright was asked to respond to Green et al. (1989) which discusses "9 years of using a three-parameter model in the construction of major achievement batteries." Here is Ben's response:

Does Green make sense? Following are some off the cuff reactions.

1. Mathematical analysis shows that the 3p model is a non-converging, inestimable elaboration of the Rasch model. When the generic criteria for measurement identified by physicists (Campbell, 1920) and mathematicians (Luce & Tukey, 1964) and demanded by the founders of psychometrics: Thorndike (1927), Thurstone (1928, 1931), Guilford (1936) and Guttman (1950a) are really required, then only the Rasch model can be deduced (Brogden, 1977; Perline, Wright & Wainer, 1979; Roskam & Jansen, 1984; Wright & Linacre, 1987; Wright, 1985, 1988a, 1988b, 1989a, 1989b). Far from being a special case of some superfluous affectation, the Rasch model is the necessary and sufficient definition of measurement. It follows that only data that can be made to fit a Rasch model can be used to construct measures.

2. Empirical analyses by Lord and Stocking demonstrate this at length:
Pages 1015-1017 (Lord, 1968) testify that: "Successful results were obtained only after a hundred or so painstaking attempts (1015)." Item discriminations "are likely to increase without limit (1015)." Person abilities "tend to increase or decrease without limit (1016)." "Divergence of the entire iterative procedure may occur simply because the initial approximations are not good enough (1016)."

Pages 13, 15 and 19 (Lord, 1975) show that even for artificial data generated to fit the 3PL model exactly only item difficulty (13) is satisfactorily recovered by LOGIST. If estimation were successful, then the dispersions of discrimination estimates (15) and guessing estimates (19) would be well estimated but, in fact, numerous of these estimates diverge by many standard errors from their generating parameter values.

Stocking (1989) compares BILOG to LOGIST unfavorably (26-28, 45) and details serious estimation problems in LOGIST (41-45). In particular: When analyzing data generated to fit the 3p model, "It is somewhat startling to find that changing starting values for item discriminations has such a large effect on the standard LOGIST procedure (24)." and "Running LOGIST to complete convergence allows too much movement away from the good starting values (25)."

More serious, "While there is no apparent bias in the ability estimates when obtained from true item parameters, the bias is significant when ability estimates are obtained from estimated item parameters. And in spite of the fact that the calibration and cross-validation samples are the same for each setting, the bias differs by test (18)." Stocking underlines this statement as well she might since it is only estimated item parameters that are available in real practice!

The startling magnitudes of bias found by Stocking are shown in her Figures 3-7 (56-60), Figures 21-23 (74-76).

Table of Contents

3. Guessing cannot and need not be estimated as an item asymptote. Guessing is inapt as an item characteristic. When guessing occurs, it is a person response anomaly, manifested occasionally by a few individuals on a few items which baffle those few persons (Wright, 1977, pp. 110-112). Only recurring lucky guessing on multiple choice items disturbs measurement. But when guesses are lucky, the consequences in the responses of the lucky guesser are clearly visible as improbable right answers. Whenever something must be done about the few lucky guesses which actually occur in multiple choice item response data, the few persons responsible for those occurrences are easy to find and reasonable corrections for any interference with measurement are easy to apply (Wright & Stone, 1979, pp. 170-190).

4. Variation in item discrimination is not only impossible to estimate without arbitrary impositions (because cross-weighing observed responses by ability estimates when discrimination is estimated and then by discrimination estimates when ability is estimated produces a regenerative feedback which escalates to infinity (Wright, 1977, pp. 103-104)) but, more devastating, modeling variation in item discrimination denies the development of construct validity because then the meaning of the variable cannot be based on item difficulty ordering. No fixed maps of item difficulty hierarchy and hence construct definition can be made because variation in discrimination forces the hierarchy of item difficult to vary with person ability. Variation in item discrimination causes ICC's to cross. But when ICC's cross, there is no unique item ordering on which to build construct validity or set standards. Construct validity and criterion meaning disappear.

5. What this means for practice is that:
a. Whenever one counts on raw scoring, i.e. counts right answers or Likert scale categories, then one is collecting data from which only a Rasch model can construct measures.
b. Whenever one estimates a regression analysis, growth study, *t*-test or means and standard deviations, one requires quantification of the dependent variable sufficiently linear and invariant to justify the arithmetic, i.e. one requires measures of the kind only Rasch models construct.
c. Whenever one aspires to understand the construct meaning of one's variables in terms of the calibrated item content by which they have been defined then one has decided to work with a model which specifies that the ICC's do not cross, i.e. a Rasch model.

6. The purpose of test analysis is not to serve the test or the variety of good and bad items which happen to fall into the test. The purpose is to serve the measurement of the child taking the test. This means:
a. Using a measurement model which establishes a clear, simple and maintainable definition of good measurement. [When one uses 3p to recalibrate the same test over samples of varying ability (an exercise any test analyzer can easily perform), the 3p estimates of discrimination and guessing are conspicuously incoherent. And even the 3p item difficulties are unnecessarily disturbed when compared with the same pair of recalibrations done by a Rasch model analysis.]
b. Using fit statistics based on this good measurement model to maintain the quality of measurement (i) by using item misfit to detect and remove eccentric items which cannot be relied upon to evoke useful responses and (ii) person misfit to identify and diagnose anomalous patterns of person response. Should some person obtain some lucky guesses, they stand out like a sore thumb against the Rasch model. [The 3p model buries this individual person information by forcing item guessing parameters on everyone who takes the items whether they guess or not.] If something beneficial, not to mention legal, is to be done about guessing, then it must face those few persons who benefit from lucky guesses and not mistreat everyone else.

Benjamin Drake Wright, 12/18/95, in a Note to Allan Olson, Northwest Evaluation Association (NWEA).

References

Green D.R., Yen W.M., Burket G.R. (1989) Experiences in the Application of Item Response Theory in Test Construction. *Applied Measurement in Education, 2*(4), 297-312.

Lord, F.M. (1968). An analysis of the Verbal Scholastic Aptitude Test Using Birnbaum's Three-Parameter Model. *Educational and Psychological Measurement, 28*, 989-1020.

Lord, F.M. (1975). Evaluation with artificial data of a procedure for estimating ability and item characteristic curve parameters. (Research Report RB-75-33). Princeton, NJ: ETS.

Stocking, M.L. (1989),. Empirical estimation errors in item response theory as a function of test properties. (Research Report RR-89-5). Princeton, NJ: ETS.

(Other references not included in the Note)

Rasch Measurement Transactions
www.rasch.org/rmt
Editor: Kenneth Royal
Email Submissions to: Editor \at/ Rasch.org

RMT Editor Emeritus: John M. Linacre
Rasch SIG Chair: Tim O’Neil
Secretary: Kirk Becker
Program Chairs: Daeryong Seo & Kelly Bradley
Rasch SIG website: www.raschsig.org

Using Rasch Measures in a Multi-level Context

These days, most educational data are conceptualized as having a nested or hierarchical structure. We frequently use test and questionnaire data for students nested within classrooms and/or schools, teachers grouped by schools, or schools contained within networks or community areas. Analysis taking into consideration the nested structure of the data enables us to partition the variance in the data to that which is between groups and the variance in the data that is between individuals within groups. Failure to partition the variance (for example, treating the between-group variance as between-individual variance) is likely to produce inferences that may not be accurate. Furthermore, using data with known measurement error enables us to separate error variance from real variance in the observations.

Any type of data can be used in hierarchical models, but Rasch measures have a particular advantage over other types of data when analyzed with this method because we can adjust for individual differences in precision. Rasch measures used at level one of a hierarchical linear model are the observed data, which contain varying amounts of measurement error. This situation results in heterogeneous variance, a possible violation of one of the basic assumptions of linear models. In equation format, the level-one relationship can be stated as:

$$Y_{ij} = \beta_{0ij} + \varepsilon_{ij}$$

Where, Y_{ij} is the outcome for individual *i* in group *j* and

$$\varepsilon_{ij} \sim N(0, \sigma_{ij}^2)$$

The error term here is not homogeneous due to the differing amounts of measurement error in the Y's. We can remove the heteroscedasticity by reweighting the measures by their inverse standard errors, or precision. So if $\hat{\sigma}_{ij}$ is the standard error of the measure Y_{ij}, we divide through by the standard error:

$$Y_{ij}^* = Y_{ij} * \left(\frac{1}{\hat{\sigma}_{ij}}\right)$$

Instead of the intercept we include $\left(\frac{1}{\hat{\sigma}_{ij}}\right)$ so

$$Y_{ij}^* = \beta_{0ij}\left(\frac{1}{\hat{\sigma}_{ij}}\right) + \varepsilon_{ij}^*$$

Where,

$$\varepsilon_{ij}^* \sim N(0,1)$$

Then β_{0ij} becomes the outcome at level two, and can be described as "the latent measure for individual *i* in group *j* adjusted for measurement error" (Raudenbush and Bryk, 2002, pp. 354-355).

This method has the immediate advantage of separating out the measurement error from the individual level error. If one were to model the outcome without the reweighting the error term would contain both the residual variation and the measurement error. In a multi-level analysis, where we are concerned about partitioning the variance into between-individuals and between-groups components, the ability to remove the error variance improves our ability to get accurate estimates of the sizes of the variances.

In addition, we can include more than one outcome, similar to multivariate regression where we can take advantage of the covariance in the outcomes to improve the prediction. Another advantage of this technique in the hierarchical context is the ability to accurately estimate group-level covariances. Multiple outcomes are included on the left side of the equation in level one, with separate indicators for each of the outcomes on the right side of the equation. So, if the outcomes are Y_{1ij} and Y_{2ij} with corresponding standard errors $\hat{\sigma}_{1ij}$ and $\hat{\sigma}_{2ij}$ the equation becomes

$$\left(\frac{1}{\hat{\sigma}_{kij}}\right)Y_{kij} = \beta_{1ij}\left(\frac{D_1}{\hat{\sigma}_{1ij}}\right) + \beta_{2ij}\left(\frac{D_2}{\hat{\sigma}_{2ij}}\right) + \left(\frac{1}{\hat{\sigma}_{kij}}\right)\varepsilon_{ij}$$

Where, $D_k, k = \{1, 2\}$ is 1 if the outcome is Y_k, 0 otherwise.

The outcomes are permitted to vary randomly at the group level, having variances and covariances described by the symmetrical matrix

$$\begin{bmatrix} \tau_{11} & \tau_{12} \\ \tau_{21} & \tau_{11} \end{bmatrix}$$

with τ_{21} giving the covariance between the two outcomes at the group level.

Weighting by the precision and treating the observations as containing differing amounts of information will probably make the estimation more efficient, and affect the sizes of the variances (both absolute and relative) and covariances. The estimation of the fixed effects is relatively robust and will probably not be affected much by the precision weighting.

In the following example, we are analyzing two survey measures, trte (Teacher-Teacher Trust) and infl (Teacher Influence). In creating the files to be analyzed in HLM, the reweighting by precision must be done externally to the program. In SAS the following lines will do the trick:

```
    if rsinfl > 0 then do;
        inflwgt = 1/rsinfl;
        trtewgt = 0;
        meas = infl/ rsinfl;
        output;
        end;
    if rstrte > 0 then do;
        inflwgt = 0;
        trtewgt = 1 / rstrte;
        meas = trte / rstrte;
        output;
        end;
```

The two measures are infl and trte. Their fit inflated standard errors[1] are rsinfl and rstrte, respectively. The equation for this model is

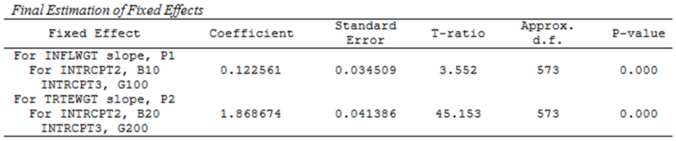
$$MEAS_{ijk} = \gamma_{100}*INFLWGT_{ijk} + \gamma_{200}*TRTEWGT_{ijk}+ r_{0jk} *INFLWGT_{ijk}+ r_{1jk} *TRTEWGT_{ijk}+ u_{10k} *INFLWGT_{ijk} + u_{20k} * TRTEWGT_{ijk}$$

In addition, in the HLM command file you have to include this statement:

```
FIXSIGMA2:1.00
```

or you will get an error message stating that there are not enough degrees of freedom available to estimate the level-1 variance.

The results for the fixed effects are:



Final Estimation of Fixed Effects

Fixed Effect	Coefficient	Standard Error	T-ratio	Approx. d.f.	P-value
For INFLWGT slope, P1					
For INTRCPT2, B10	0.122561	0.034509	3.552	573	0.000
INTRCPT3, G100					
For TRTEWGT slope, P2					
For INTRCPT2, B20	1.868674	0.041386	45.153	573	0.000
INTRCPT3, G200					

The fixed-effect estimates for infl and trte from the model without a measurement model at level one are 0.1246 and 2.0518, respectively. In general, the fixed-effect estimates are quite robust and will not be substantially affected by the addition of the measurement model. The random effects are a different story. Here is the level-3 variance-covariance matrix from the model with the measurement model at level one.

```
tau(beta)
INFLWGT                  TRTEWGT
INTRCPT2,B10             INTRCPT2,B20
0.57039                  0.39288
0.39288                  0.72401
```

Here is the corresponding matrix from the model without a measurement model:

```
tau
INFLIND,B1    0.75137     0.59118
TRTEIND,B2    0.59118     0.93246
```

The differences are quite substantial. As you would expect, adjusting for the measurement error in the observations reduces the size of the group-level variances. Moreover, the intra-class correlation for the model with the measurement model is 0.32 and 0.19, for infl and trte, respectively. This indicates that 32 percent of the variance in infl is between schools, and the remainder is within schools, among teachers. The corresponding ICCs for the model without the measurement model are 0.17 and 0.20. However, which of these estimates is closer to the truth is an open question. Surely adjusting for different amounts of information in the observations will result in more efficient estimation. This is analogous to using GLS instead of OLS when the assumption of homogeneous variance does not hold. So, in theory, we can say that if you know the amount of error in each of your measurements you might as well take advantage of this knowledge in your analyses. But in practical terms, exactly how much your estimates will differ is not clear. I am presently working on a simulation study that will determine the concrete effects on the fixed and random effect estimates of the addition of a measurement model at level one of the HLM.

Reference

Raudenbush, S.W. and Bryk, A.S. (2002). *Hierarchical Linear Models* (Second Edition). Thousand Oaks: Sage Publications.

[1] The fit inflated standard error is $rse = se * (\max(1, inmnsq, outmnsq))^{1/2}$ where se is the model standard error. Sometimes we include 1/inmnsq and 1/outmnsq inside the max() but especially with measures constructed from survey data where respondents may skip any items they do not feel like answering, there can be tremendous overfit, which overweights the inflated standard errors.

Stuart Luppescu
University of Chicago Consortium on Chicago School Research

Call for Submissions

Research notes, news, commentaries, tutorials and other submissions in line with *RMT*'s mission are welcome for publication consideration. All submissions need to be short and concise (approximately 400 words with a table, or 500 words without a table or graphic). The next issue of *RMT* is targeted for Dec. 1, 2013, so please make your submission by Nov. 1, 2013 for full consideration. Please email Editor\at/Rasch.org with your submissions and/or ideas for future content.

Disconnected Subsets, Guttman Patterns and Data Connectivity

Rasch methodology implements conjoint measurement. Ideally, all the Rasch parameters (person, items, raters, tasks, rating-scale thresholds, etc.,) are placed in one measurement frame-of-reference so that the estimate of each parameter is located unambiguously relative to the estimate of every other parameter. Sadly, empirical data often fail to support this ideal. The most frequently encountered failures are extreme scores. If a person succeeds on every item on a standard multiple-choice test, then that person obtains the maximum possible score, 100%, and the Rasch estimate corresponding to that score is infinity. In practice, a finite, but outlying, estimate is reported for Rasch measure corresponding to the extreme score (Wright, 1998). Other failures are fortunately rarer.

Disconnected Subsets

These can be encountered in judge-intermediated data but they sometimes also occur in adaptive or tailored tests and surveys. Table 1 is a simple example of a dichotomous dataset with disconnected subsets.

Table 1. Disconnected Subsets

	Item 1	*Item 2*	*Item 3*	*Item 4*
Person A	0	1	*m*	*m*
Person B	1	0	*m*	*m*
Person C	*m*	*m*	1	0
Person D	*m*	*m*	0	1

m = missing data, not administered

Persons A and B both scored 1 on Items 1 and 2, so their estimated Rasch ability measures are the same. Persons C and D both scored 1 on Items 3 and 4, so their estimated Rasch ability measures are the same. But how do the estimates for Persons A and B relate to the estimates for Persons C and D? At first glance, they all scored 1 so their estimates are all the same, but this assumes that Items 3 and 4 have the same difficulty as Items 1 and 2. What if Items 3 and 4 were more difficult than Items 1 and 2? Then Persons C and D scored 1 on more difficult items, and so their estimated abilities would be higher than the estimates for Persons A and B. Or, what if Items 3 and 4 were easier? Then Persons C and D would have lower estimates. We see that Persons A and B with Items 1 and 2 are one subset of the data. Persons C and D with Items 3 and 4 are another subset of the data. Estimates of the parameters in one of the subsets cannot be compared unambiguously with estimates of the parameters in the other subset. The disjoint subsets of data are in different frames-of-reference.

Disconnected subsets are not always obvious in rater-intermediated data. The judging plan may specify that each examinee is rated by a pair of raters, and that the pairs of raters change partners according to the judging plan at the start of each judging session. However, unless the raters are carefully supervised, they may not follow the plan. At worst, they may not change partners at all! If this happens, pairs or groups of raters may bring about disconnected subsets of ratings in the data. All the examinees may be rated on the same items, but there are subsets of raters and examinees with no overlap with other subsets of raters and examinees. Accordingly, it is vital to start data analysis as soon as the first ratings are collected so that problems in the operation of the judging plan can be quickly identified and remedied before the judging process has been completed.

If disconnected subsets in the data are not identified until after data collection has completed, then constraints must be imposed on the Rasch measures in order to make them approximately comparable. For instance, in a judging situation, we may say that the mean abilities of the examinees in each subset are the same, because the examinees were assigned to judges at random. Alternatively we might say that the mean leniency of the subsets of judges is the same because the judges were assigned initially at random and they had all participated in the same training sessions. However, these constraints inevitably have an arbitrary aspect to them. Some examinees will be advantaged and some disadvantaged. As Shavelson and Webb (1991) remark, it is “the luck of the draw”.

Guttman Patterns

Psychometrician Louis Guttman (1916-1987) perceived the ideal test to be one in which a person succeeds on all the items up to a certain difficulty, and then fails on all the items above that difficulty. Then, when persons and items are ordered by raw score, this produces a data set with a “Guttman pattern”. A Guttman pattern is shown in Table 2.

Table 2. Guttman Pattern

	Item 1	*Item 2*	*Item 3*	*Item 4*	*Person score*
Person A	1	1	1	1	4
Person B	1	1	1	0	3
Person C	1	1	0	0	2
Person D	1	0	0	0	1
Item score	4	3	2	1	

These data are very orderly. Person A performed better than Person B, who performed better than Person C, who performed better than person D. But what about measuring the performances? Is the difference between Person A and Person B greater or less than the difference between Person C and Person D? Figure 1 shows two depictions of an additive conjoint latent variable. For both of them, the most likely data is the Guttman pattern in Table 2. There is no information in the data about which of these depiction is more accurate. Georg Rasch perceived that there must be probabilistic disordering ("Guttman reversals") in the data in order to quantify the distance between two elements (persons, items, raters, etc.). A more able person must fail on an easier item, or a less able person must succeed on a more difficult item in order for the distances between the persons to be additively quantifiable.

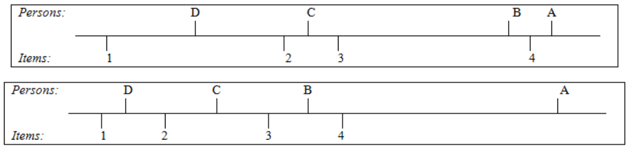


Figure 1. Two depictions of a latent variable that accord with the Guttman Pattern in Table 2.

Guttman Splits

Guttman patterns are rarely observed in empirical datasets. The Guttman *Coefficient of Reproducibility* is rarely 1.0, but there can be unnoticed Guttman Splits in the data. Table 3 illustrates this. In Table 3, every person and item appear to be estimable, because every row and every column has some successes (1) and some failures (0). There are no extreme scores for persons or items. We see that Persons A and B are more able than Persons C and D, also that Items 3 and 4 are more difficult than Items 1 and 2. However, there is a Guttman split between Persons B and C, and between Items 2 and 3. There is no item in the data where Persons A or B fail and Person C or D succeed. Also there is no person in the data for whom there is successs on Items 3 or 4 and failure on items 1 or 2. Persons A, B and Items 3, 4 are all at one location on the latent variable. Also, Persons C, D and Items 1, 2 are all at another location on the latent variable. Regretably, there is no information in the data for estimating the distance between those two locations.

Table 3. Guttman Split

	Item 1	*Item 2*	*Item 3*	*Item 4*	*Person score*
Person A	1	1	0	1	3
Person B	1	1	1	0	3
Person C	0	1	0	0	1
Person D	1	0	0	0	1
Item score	3	3	1	1	

A Practical Example of a Guttman Split

An Olympic Ice-Skating dataset, Exam15.txt in the Winsteps Examples folder, has been analyzed many times. Its estimates are slow to converge, requiring more than 700 iterations through the data, depending on the convergence criteria, much more than the 20 iterations or so required for most datasets. The reason for the slowness in estimation is that there is a Guttman Split in the dataset (which I did not notice for ten years). This is shown in Table 4. Each Judge gave each Skating Performance a score in the range 0.0 to 6.0. These are analyzed as ratings on a scale from 0 to 60. Performance Numbers 1 to 5 all received ratings of 58 and 59. The highest rating given to any of the other 75 Performances is 58. There is a Guttman Split between Performances 5 and 6. We know that the top 5 Performances are better than the other 75 performances, but the data do not tell us how much better in Rasch terms.

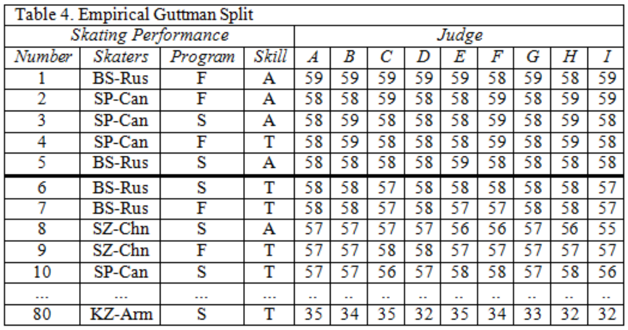


Table 4. Empirical Guttman Split

Skating Performance				*Judge*								
Number	*Skaters*	*Program*	*Skill*	*A*	*B*	*C*	*D*	*E*	*F*	*G*	*H*	*I*
1	BS-Rus	F	A	59	59	59	59	59	58	59	58	59
2	SP-Can	F	A	58	58	59	58	58	59	58	59	59
3	SP-Can	S	A	58	59	58	58	58	59	58	59	58
4	SP-Can	F	T	58	59	58	58	58	59	58	59	58
5	BS-Rus	S	A	58	58	58	58	59	58	58	58	58
6	BS-Rus	S	T	58	58	57	58	58	58	58	58	57
7	BS-Rus	F	T	58	58	57	58	57	57	58	58	57
8	SZ-Chn	S	A	57	57	57	57	56	56	57	56	55
9	SZ-Chn	F	T	57	57	58	58	57	57	57	57	57
10	SP-Can	S	T	57	57	56	57	58	58	57	58	56
...	...	...	...	..	..	..	..	..	..	..	..	..
80	KZ-Arm	S	T	35	34	35	32	35	34	33	32	32

Workarounds for Disjoint Datasets and Guttman Splits

The best solution to this type of problem is to analyze the data as they are being collected. Then problems in the data can be identified and remedial action taken before data collection has finished. For instance, the judging plan can be adjusted or extra data can be collected. After data collection has finished, there are two approaches:

(1) Add reasonable dummy data records to the dataset to produce reasonable estimates. The parameters (persons, item, thresholds, etc.) can then be anchored at their reasonable values and the dummy data records omitted for the final reporting. In Table 4, we could add andditional dummy Judge J who gives Performance 5 a rating of 57 and Performance 6 a rating of 58. Now all the Performances can be estimated uniquely in one frame of reference. After anchoring, the dummy Judge would be omitted for the final reporting.

(2) Put reasonable constraints on the estimates. For instance, in Table 4, we might decide that Performance 5 is one logit better than Performance 6. According, Performance 5 is anchored (fixed) at +1.0 logits and Performance 6 at 0.0 logits. The Performances can now be estimated uniquely in one frame of reference. For disconnected subsets, such as Table 1, reasonable constraints may be that the mean ability of the two subsets of persons is the same or the mean difficulty of the two sets of items is

the same. Alternatively, the items might be aligned on the latent variable using Virtual Equating (Luppescu, 2005).

References

Luppescu S. (2005). Virtual Equating. *Rasch Measurement Transactions, 19*(3), p. 1025. www.rasch.org/rmt/rmt193a.htm

Shavelson, R. J., & Webb, N. M. (1991). *Generalizability Theory: A Primer*. Thousand Oaks, CA: Sage.

Wright B.D. (1998). Estimating Rasch measures for extreme scores. *Rasch Measurement Transactions, 12*(2), p. 632-3. www.rasch.org/rmt/rmt122h.htm

John Michael Linacre

A Cautionary Tale about Item Equating with Fluctuating Samples

In many high-stakes testing scenarios samples tend to be reasonably comparable with regard to demographic characteristics across administrations. As part of the initial quality control checks, most psychometricians will investigate various demographic characteristics to get a pulse on sample stability. Unfortunately, many psychometricians may be tempted to only investigate "visible" demographic variables, such as gender, ethnicity, and so on. Failing to investigate "invisible" demographic variables such as whether the examinee is a first-time or repeat test-taker, or has previously rendered a fail result could lead to an enormous mistake with regard to equating examinations. Consider the following example.

Suppose a data set is provided to a psychometrician for scoring. As part of the initial quality control checks, s/he learns both the sample size and the visible demographics variables all seem fairly comparable to previous administrations. On the surface it appears the sample is comparable to previous samples, thus the psychometrician proceeds to investigate item quality and functioning. Preliminary item analyses reveal the items appear to be sound and functioning properly. Upon obtaining this assurance, the psychometrician then begins developing item anchors for equating purposes. After several iterations of investigating displacement values and unanchoring item calibrations that displace from those obtained from previous administrations, the psychometrician is satisfied with the remaining item calibrations and locks them down as anchors for the final scoring run.

Once data are scored, the results are reviewed and compared to historical trends. Diagnostic results (e.g., fit statistics, separation and reliability estimates, etc.) appear sound, but some notable differences in pass/fail statistics and mean scaled scores are evident. Concerned, the psychometrician revisits the scoring processes by reviewing syntax and reproducing all relevant data files. Examination data are rescored and the same results are produced. Still suspicious, the psychometrician begins combing both the new data set and last year's data set to identify anyone that had previously taken the exam. A list of repeat examinees is pulled and their scores are compared across both administrations of the examination. It turns out virtually all of the repeat examinees appear to have performed worse on the new examination. How could this be? Examinees have had additional training, education and time to prepare for the examination.

Upon closer inspection the psychometrician is surprised to learn a less obvious demographic characteristic had fluctuated among the examinees and caused this unusual scenario. It turns out a larger proportion of examinees were taking the examination due to a prior failure. This small, yet very important, artifact had a significant ripple effect on the quality of the final scores. The problem began when a less able sample interacted with items and the psychometrician was deceived into thinking many of the existing calibrations were unstable. As a result, the psychometrician unanchored many item calibrations that should otherwise have been left alone. Thus, when the new scale was established, it jumped and resulted in scores that lost their meaning across administrations.

Although item equating under the Rasch framework is quite simple and straight-forward, it still requires a great deal of careful attention. The scenario presented above illustrates how a significant problem may occur simply as a result of failing to investigate one key demographic characteristic of the sample. When equating, it is critical that one considers all types of sample characteristics, especially those that pertain to previous performance. An inconsistency in these demographics can result in item instability, which in turn, can go unnoticed when examining displacement values and creating item anchors. It is for this reason that many psychometricians only use first-time examinee data when equating exams. In any instance, all psychometricians that equate examinations under the Rasch framework would be wise to include to their list of quality control checks a comprehensive investigation of demographic characteristics both before and after a scoring run is complete.

Kenneth D. Royal, University of North Carolina at Chapel Hill
Mikaela M. Raddatz, American Board of Physical Medicine and Rehabilitation

Rasch Measurement in the News

A recent article in *Education Week* discussed the potential wide-scale application of measures produced from Rasch models. The author of the article, Tom Vander Ark, cites the enormous mounds of fragmented, educational data currently available and the lack of a common scale for meaningful reporting and growth measurement. Vander Ark proposes using the widely used Lexile framework for reading and the Quintile framework for math to link scales and develop a system of comparable growth measures for students. Metametrics' Gary Williamson was quoted as saying "For the best measurement of student growth, the measurement scale must be: unidimensional, continuous, equal-interval, developmental, and invariant with respect to location and unit size. In fact, the Lexile scale possesses all of these necessary characteristics. So it is not only appropriate for the measurement of student growth, it may well be the most appropriate scale for the measurement of academic growth in reading".

Vander Ark, T. (2013). A proposal for better growth measures. *Education Week*. June 5. Available at: http://blogs.edweek.org/edweek/on_innovation/2013/06/a_proposal_for_better_growth_measures.html

Measurement and Assessment in Higher Education

NEW Special Interest Group (SIG) within the American Educational Research Association

We would like to introduce a new scholarly community within the American Education Research Association (AERA). As the name implies, the *Measurement and Assessment in Higher Education* SIG focuses on measurement and assessment issues in higher education, specifically those related to student learning outcomes. We emphasize (1) the methodological challenges encountered in assessment practices for program improvement and/or accountability purposes (e.g. sampling, psychometrics, validity, innovative item types, performance assessment, missing data) and (2) the strategies to address them. If interested, we encourage you to attend this SIG's sessions at the AERA annual meeting in Philadelphia, which will be held April 3 to April 7, 2014.

If you have any questions about submitting a proposal for the 2015 conference or would like to know more about this SIG, please do not hesitate to contact the program chair (Katie Busby, kbusby~at~tulane.edu) or the SIG chair (Keston Fulcher, fulchekh~at~jmu.edu) who will be in attendance at the 2014 AERA conference.

Building measures from raw scores – We need to use the Wright stuff!

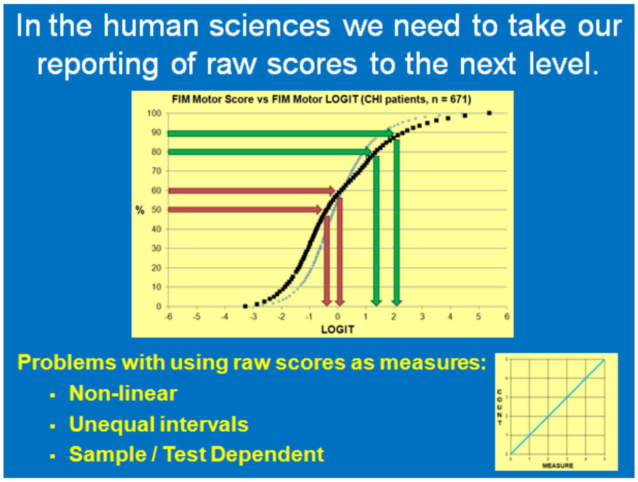
This is a short exercise designed to get people thinking about the difference between counting and measuring. This exercise utilizes the fact that balloons are highly similar manufactured products. But they can be manipulated by the trainer in terms of their number, size and color. Helium filled balloons are also easy to obtain and visually engaging. They are large enough for students to see in any classroom. This presentation is dedicated to Dr. Benjamin D. Wright - psychological measurement hero! It starts with the replication of some famous graphs by Dr. Wright (Wright 1993, Wright 1997).

"As researchers and clinicians in the human sciences we need to take our reporting of raw scores to the next level. Raw scores have three fundamental problems when they are used as measures. They are: non-linear, based on unequal intervals, and sample or test dependent.

The graph in Figure 1 shows the relationship between these raw score percentages with their corresponding linear measure. It shows:

- The non-linearity of raw scores, especially at the extremes of a scale.
- That a 10 percent change in raw scores does not produce the same result across a scale. It depends on where you start.
- That raw scores derived from different scales have a different relationship to the common linear measure."

Figure 1: Summary slide showing FIM Motor Raw Scores versus FIM Motor LOGITs (Closed Head Injury patients, n = 671)



In the human sciences we need to take our reporting of raw scores to the next level.

Problems with using raw scores as measures:

- Non-linear
- Unequal intervals
- Sample / Test Dependent

"I would now like to demonstrate these points with some simple exercises. They highlight the difference between counting and measuring, and introduce the concept of the number line."

*The correct answers to my questions are in **bold**.*

DEMONSTRATING NON-LINEARITY

Show 3 helium filled balloons of the same size and color (NB: each balloon is tied to a ribbon). Holding the three balloons in one hand, ask the following question:

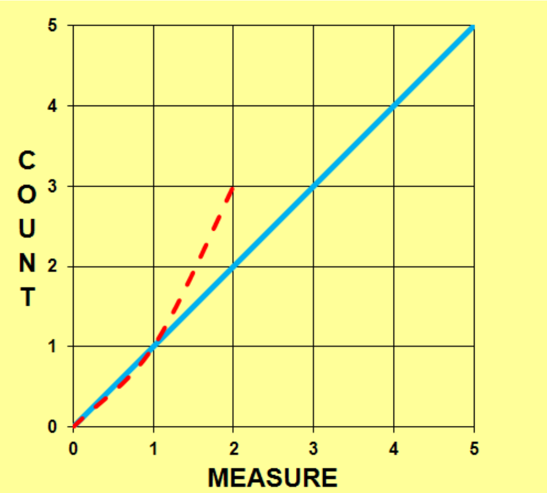
"How many balloons am I holding?" **3**

Now release some of the helium from one of the three balloons. Hold them up again and ask:

"How many balloons am I holding now?" **3**

Then ask "Which set of balloons would you prefer to have?" **The first set**

"Can you see the difference? In one set of balloons we are counting and in the other set we are measuring." This example can be seen on the number line below:



This example shows the difference between counting and measuring, and the need to have a one to one correspondence between our counts and our measures.

DEMONSTRATING UNEQUAL INTERVALS

Get a set of 3 balloons, each tied to a weight. Use one large balloon followed by one small balloon and then one large balloon. Place them on the left hand side of the room. Space them about one meter apart.

Then get another set of 3 balloons each tied to a weight. This time the balloons are of equal size. Place them on the right hand side of the room. Space them about one meter apart.

Stand behind the balloons on the left hand side of the room. Start with the balloon on your right.

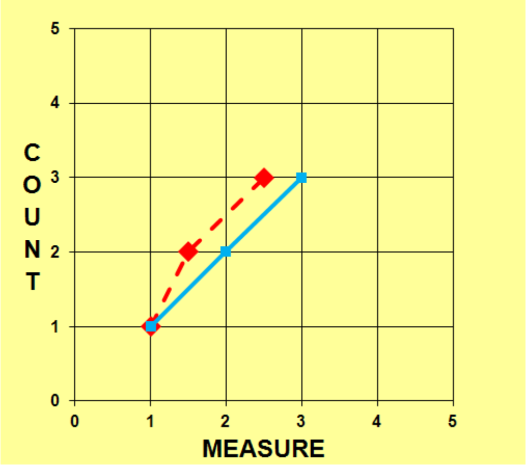
"I start from here (*Step across the next two balloons*) and move to here, taking the balloons as I go."

Now go to the example on the right hand side of the room. Stand behind the balloons and start with the balloon on your right.

"I start from here (*Step across the next two balloons*) and move to here, taking the balloons as I go."

"Now, in which example have I made the most improvement or gain?" **The second example**

This example can be seen on the number line below:



"Now imagine that I painted the numbers one, two and three on the front of these balloons. Would this make any difference to your answer?"
(This comment brings up the issue of using numbers as object labels or identifiers.)

"Here is another example."
Show a set of 3 balloons: one small, one medium and one large.

"Here the balloons are in size order. The last balloon is bigger than the middle balloon, which is bigger than the first balloon."
Grab the smallest and the medium balloon. Hold them in one hand and hold the largest balloon in your other hand.

"Now, I have these two balloons. If I put these two balloons together (i.e. combine them) are these two balloons bigger than or smaller than this balloon?"
Shake the largest balloon in your other hand.
Smaller / About the same / Bigger / Not sure

"By how much?" **Don't Know**

"Now, what if I had different sized balloons even though they were in the same order would I get a different answer?" ***Yes***

(It depends on the sizes of the two balloons chosen)
Now get a set of 3 balloons of equal size. Hold them in one hand.

"These balloons are the same size. Now I grab these two balloons."
Hold them in one hand and hold the other balloon in your other hand.

"If I put these two balloons together, combining them, are these two balloons bigger than or smaller than this balloon?"
Shake the balloon in your other hand. **Bigger**

"By how much?" **1 balloon**

"Was that difference much easier to work out?" ***Yes***

"Knowing the amount or quantity (or the difference between amounts / quantities - i.e. how much more or less) is the key question for measurement."
These two examples show why we need to know the intervals between our raw scores. Having equal intervals or units is also important for interpretation.(These two examples can be used as a starting point for discussions about: idealized measurement, the concepts of quantity and number, conjoint measurement, logarithmic data transformations, ogives, concatenation, frames of reference, exchangeable units, rulers and other measurement analogies – for example, balance scales, ammeters and thermometers.)

DEMONSTRATING TEST / SAMPLE DEPENDENCY

Show another set of 3 balloons: one blue, one red, one green.

"Finally, I have a couple more questions to ask."
Question 1: How many balloons am I holding? **3**

Question 2: How many balloons am I holding which are a primary color for painting? **2**

Can you see that the sample has not changed only the question has changed?"
This example illustrates how our raw scores are based on the sample tested or the questions asked. (This example can then be used to generate further discussions about stochastic measurement and latent traits.)

"In conclusion, these examples have shown the problems with raw scores – their non-linearity, their unequal intervals, and their sample / test dependence. These problems emphasize the difference between counting and measuring. When reporting our results in the human sciences we must go beyond counting the 'correct answers' to our questions or counting the 'ordered numbers' from our rating scales. We can start to solve these issues when we think about the measurement of our variables in terms of a number line with equal interval characteristics."
Working with helium balloons is fun, though it requires a bit of practice. (I would also recommend using a classroom with low to normal ceiling height.)

References

Linacre, J. M., & Wright, B. D. (1993). Constructing linear measures from counts of qualitative observations. Paper presented at the *Fourth International Conference on Bibliometrics, Informetrics and Scientometrics*, Berlin, Germany, September 1993 (ERIC ED 364574).

Merbitz, C., Morris, J., & Grip, J. C. (1989). Ordinal scales and foundations of misinference. *Archives of Physical Medicine and Rehabilitation, 70*, 308-312.

Silverstein, B., Kilgore, K., & Fisher, W. (1989). Don't abandon FAS (Letter to the Editor in response to Merbitz, Morris & Grip 1989). *Archives of Physical Medicine and Rehabilitation, 70*, 864-865.

Wright, B. D. (1993). Thinking with raw scores. *Rasch Measurement Transactions, 7*(2), 299-300.

Wright, B. D. (1997). Fundamental measurement for outcome evaluation. *Physical Medicine and Rehabilitation: State of the Art Reviews, 11*(2), 261-288.

Wright, B. D., & Linacre, J. M. (1989). Observations are always ordinal: Measurements, however, must be interval. *Archives of Physical Medicine and Rehabilitation, 70*, 857-860.

Wright, B. D., & Mok, M. M. C. (2004). An overview of the family of Rasch Measurement Models. In E. V. Smith Jr. & R. M. Smith (Eds.), *Introduction to Rasch measurement*. Maple Grove, MN: JAM Press.

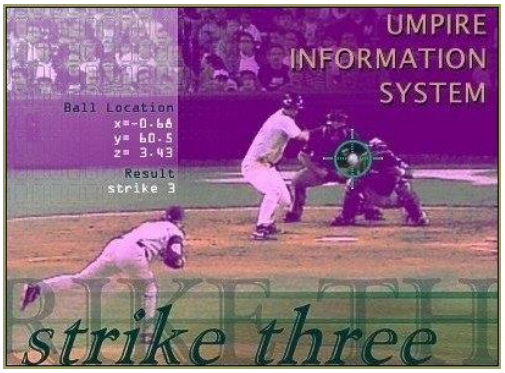
Nick Marosszeky
Macquarie University, Australia

Kaggle Competitions

Rasch measurement enthusiasts may be interested in participating in one or more of Kaggle's open competitions. The Kaggle website features a steady stream of data science competitions. Some notable competitions currently underway include the "Recognize users of mobile devices from accelerometer data", "Personalize Expedia hotel searches-ICDM 2013", and the "Belkin energy disaggregation competition". Visit www.kaggle.com for a comprehensive list of current competitions, descriptions of the competitions, and prize monies.

Measurement in Sports: Umpire Information System

QuesTec, a leader in measurement technologies, has teamed with Major League Baseball to maintain the Umpire Information System (UIS). According to QuesTec, "The UIS uses QuesTec's proprietary measurement technology that analyzes video from cameras mounted in the rafters of each ballpark to precisely locate the ball throughout the pitch corridor. This information is then used to measure the speed, placement, and curvature of the pitch along its entire path. The UIS tracking system is a fully automated process that does not require changes to the ball, the field of play, or any other aspect of the game. Additional cameras are mounted at the field level to measure the strike zone for each individual batter, for each individual pitch, for each at bat. This information is compiled on a CD ROM disk and given to the home plate umpire immediately following each game."



How does it work?

"QuesTec technology actually measures information about interesting events during the game that would not be available any other way… The ball tracking component uses cameras mounted in the stands off the first and third base lines to follow the ball as it leaves the pitcher's hand until it crosses the plate. Along the way, multiple track points are measured to precisely locate the ball in space and time. This information is then used to measure the speed, placement, and curvature of the pitch along its entire path. The entire process is fully automatic including detection of the start of the pitch, tracking of the ball, location computations, and identification of non-baseball objects such as birds or wind swept debris moving through the field of view. No changes are made to the ball, the field of play, or any other aspect of the game, to work with QuesTec technology." QuesTec claims the technology is accurate to within 0.5 inch.

QuesTec Umpire Information System. Available at: http://www.questec.com/q2001/prod_uis.htm

Journal of Survey Statistics and Methodology

Some Rasch measurement enthusiasts may be interested in a new journal devoted to survey research. The *Journal of Survey Statistics and Methodology*, sponsored by AAPOR and the American Statistical Association, began publishing in 2013.

The journal will publish both theoretical and applied papers, provided the theory is motivated by an important applied problem and the applied papers report on research that contributes generalizable knowledge to the field. Review papers are also welcomed. Papers on a broad range of surveys are encouraged. The journal will contain three sections: 1) Survey Statistics; 2) Survey Methodology; and 3) Applications. For more information about the journal please see http://jssam.oxfordjournals.org/.

Journal of Applied Measurement Vol. 14, No. 3, 2013

Richard M. Smith, Editor, www.jampress.org

Rasch-related Coming Events

Sept. 13-Oct. 11, 2013, Fri.-Fri. Online workshop: Rasch Applications in Clinical Assessment, Survey Research, and Educational Measurement (W. P. Fisher), www.statistics.com

Sept. 18-20, 2013, Wed.-Fri. In-person workshop: Introductory Rasch (A. Tennant, RUMM), Leeds, UK, www.leeds.ac.uk/medicine/rehabmed/psychometric

Sept. 23-25, 2013, Mon.-Wed. In-person workshop: Intermediate Rasch (A. Tennant, RUMM), Leeds, UK,

Sept. 26-27, 2013, Thur.-Fri. In-person workshop: Advanced Rasch (A. Tennant, RUMM), Leeds, UK

Oct. 2, 2013, Wed. Online free webinar: Demystifying Rasch Analyses for Clinical Application: Item Response Theory in Clinical Practice (J. Moore, C.H. Chang).

Oct. 18- Nov. 15, 2013, Fri.-Fri. Online workshop: Practical Rasch Measurement – Core Topics (E. Smith, Winsteps), www.statistics.com

Oct. 20 – Oct. 25, 2013, Sun. – Fri. International Association for Educational Assessment (IAEA) 39th Annual Conference, Tel Aviv, Israel, www.iaea-2013.com

Oct. 31 – Nov. 2, 2013, Thurs.-Sat. In-person workshop: Rasch Measurement (R. Smith, N. Bezruczko, S. Wind, IPARM, Winsteps), Maple Grove, MN, www.jampress.org/workshops.htm

Nov. 22, 2013, Fri. 8th Workshop on Rasch models in business research, La Laguna, Tenerife, www.insitutos.ull.es/viewcontent/institutos/iude/46416/es

Dec. 11-13, 2013, Wed.-Fri. In-person workshop: Introductory Rasch (A. Tennant, RUMM), Leeds, UK,

Jan. 3-31, 2014, Fri.-Fri. Online workshop: Practical Rasch Measurement – Core Topics (E. Smith, Winsteps), www.statistics.com

Mar. 12-14, 2014, Wed.-Fri. In-person workshop: Introductory Rasch (A. Tennant, RUMM), Leeds, UK,

Mar. 31-Apr. 1, 2014, Mon.-Wed. IOMW Biennial Meeting. Philadelphia, PA, www.iomw2014.eventbrite.com

Apr. 3-7, 2014, Thurs.-Mon. AERA Annual Meeting, Philadelphia, PA, www.aera.net