*Metodología de las Ciencias del Comportamiento 5(2)*, 179–194 

# Validity and assessment: a Rasch measurement perspective

Trevor G. Bond [1]

*School of Education, James Cook University Q 4811, AUSTRALIA*

**Abstract**

This paper argues that the Rasch model, unlike the other models generally referred to as IRT models, and those that fall into the tradition of True Score models, encompasses a set of rigorous prescriptions for what scientific measurement would be like if it were to be achieved in the social sciences. As a direct consequence, the Rasch measurement approach to the construction and monitoring of variables is sensitive to the issues raised in Messick's (1995) broader conception of construct validity. The theory / practice dialectic (Bond & Fox, 2001) ensures that validity is foremost in the mind of those developing measures and that genuine scientific measurement is foremost in the minds of those who seek valid outcomes from assessment. Failures of invariance, such as those referred to as DIF, should alert researchers to the need to modify assessment procedures or the substantive theory under investigation, or both.

KEY WORDS: *Construct validity, Rasch, IRT, measurement*

## Whose validity is more important?

"Yes, I think I do realize what you're saying to me. You're saying that item 17 on the test is such an important question about the ability being tested that you can't possibly do away with it. However, what I'm trying to say you is, that, in spite of your attempts so far, it is highly unlikely that the ability that the examinees use to respond correctly to item 17 is the same sort of ability that underlies their success on the other items in the test."

"How can you say that? The item was written by the same members of the examiners board who wrote the other items; they think that item is crucial in terms of the ability being tested. In fact, you can't take it out without damaging the validity of the test. We're the examining board; item 17 has to stay."

"The problem, basically, is this. You intend to grade people on this test according to the well used practice of summing the total scores. There's nothing wrong with that - in fact it's a great idea, but success on item 17 does not contribute to that total score in the same way as success on the other items does. Examinees who have high scores on the other items sometimes fail on #17; some of the lowest ability children are somehow getting this item right. At this stage item 17 is not communicating what you want to test to the examinees sufficiently well that their responses to that item should be counted in the test score over all. The safest thing for us to do at this

[1]Dirección para correspondencia: Trevor G Bond, Ph D, School of Education, James Cook University Q 4811. AUSTRALIA. Email: <trevor.bond@jcu.edu.au>

stage is to omit this item from the total score. Then, we should seek out some of the children who performed unexpectedly and to try to find out how your assessment intentions went wrong. After that we can modify item 17 or we might write a new item for inclusion in a further test."

"We've already included #17 in the test – we're not going to throw the results away just because you don't seem to like them."

At least, in the case of educational testing, stories like this one are legion. The team of content experts and the psychometrician go head to head on an important testing issue. Often, it is face validity, as judged by the expert panel, that wins the day - with the psychometrician walking out, mumbling, "Fine. Just don't add the scores together!"

The aim of this paper is to address issues of construct validity in psychological and educational assessment from the perspective provided by Rasch measurement. This, of course, focuses on only one small part of the sort of research routinely carried out in the social sciences. While a considerable amount of this research takes an unabashedly qualitative view point, the issues here pertain directly to those who adopt a quantitative approach to their research in the human sciences. And, from within that quantitative perspective, it addresses those who attend to the issue of the measurement of human variables and, in particular, those who attempt to emulate in their research some of those particular properties of measurement that abound in the physical sciences.

## The central role of measurement

The Rasch model, unlike the other models generally referred to as IRT models, and those that fall into the tradition of True Score models, encompasses a set of rigorous prescriptions for what scientific measurement would be like if it were to be achieved in the social sciences. Those measurement prescriptions might be generally described as a probabilistic form of the axiomatic conjoint measurement recommended to us by a Duncan Luce et al. (Luce & Tukey, 1964) in the first instance and Joel Michell and others thereafter. In a very recent article, "Measurement: A Beginner's Guide", Michell (2003) offered:

> Some other theories, such as Rasch's item response model (Rasch, 1960), are amenable to the application of conjoint measurement. According to Rasch's model, the relationship between the probability of a person, $i$, performing correctly on a unidimensional, dichotomous test item, $j$, $P(\chi^2_{ij} = 1)$; $i$'s level of the relevant ability, $a_i$,; and $j$'s level of difficulty, $d_j$, is
>
> $$P(\chi^2_{ij} = 1) = \frac{e^{(a_i - d_j)}}{1 + e^{(a_i + d_j)}} \quad (1)$$
>
> (where $e$ is the base of the natural logarithm function). If (1) is true, then
>
> $$P(\chi^2_{ij} = 1) = f(a_i - d_j) \quad (2)$$

> ($f$ is an increasing monotonic function). That is, if Rasch's theory is true, then for any person, $i$, and item, $j$, of the relevant kind, the probability of $i$ getting $j$ correct increases with the difference between $i$'s ability and $j$'s difficulty. Proposition (2) implies that order relations between these probabilities across persons and items must satisfy the hierarchy of conjoint measurement cancellation conditions (Michell, 1990)."
>
> (Michell, 2003, p.306)

While it can be readily demonstrated that the matrix of *expected* response probabilities table derived from any set of Rasch item and person estimates will satisfy those key measurement axioms, it is equally obvious that those axioms are routinely violated in the 2-PL and 3-PL IRT models (Karabatsos, 1999a, b; 2000). It is the so far unresolved challenge for Rasch measurement to demonstrate that the procedures for determining whether the matrix of *actual* response frequencies adheres sufficiently to the Rasch measurement prescriptions to satisfy the key conjoint measurement axioms. If it is the case that the practice of Rasch measurement does not yet satisfy this requirement, it might be allowed that this is because those interested in the development of the genuinely interval linear scale measures in the social sciences have set their standards appropriately high.

Elsewhere, I have argued (Bond & Fox, 2001) that attention to stringent measurement requirements in the social sciences goes hand-in-hand with attention to the all-encompassing concept of construct validity as expounded by Samuel Messick over the last decade. While Messick's thesis is that his extended conception of construct validity is equally relevant to both quantitative and qualitative psychological assessments, its use in the current context refers to those quantitative summaries of assessments that we routinely call 'test scores'. Messick further asserted that, "The extent to which score meaning and action implications hold across persons or population groups and across settings or contexts is a persistent and perennial empirical question."(1995, p.741)

Our attention to this aspect of construct validity, should derive, I contend, from the principle of *invariance* of Rasch measures: estimates of item difficulty on the one hand and estimates of person ability on the other. Given that Rasch measurement instantiates interval level, rather than ratio level measurement, invariance of item and person estimate values remains relative. Individual Rasch analyses (by default) adopt the mean of the item difficulty estimates as the zero point on the calibration scale, so it is the differences between item and person estimates, rather than those estimates per se which should remain invariant across investigations.

It is a *prima facie* requirement of measurement outside the social sciences that the values attributed to variables by any measurement system should be independent of the particular measurement instrument used (as long as the instrument is appropriate to the purpose). Moreover, the calibrations of the measurement instrument should remain invariant across any of its intended purposes. Given our relatively brief history of attempting genuine measurement in the human sciences, it must be expected that our measurement attempts should fall somewhat short of the physical sciences model. However, I would argue, only the fainthearted or those with low expectations will yield to the obvious difficulties that must be overcome if measurement rather than mere quantification is going to be the hallmark of our research.

An important goal of early research should be the establishment of item difficulty values for important testing devices along with the anchoring of further investigations to the item values derived from that earlier research. It would be naïve however to expect our early attempts will satisfy even the relative invariance principle on all appropriate occasions. In the same way as the Inspector of Weights and Measures would monitor the scales used by the local greengrocer, we would continually monitor that the item values continued to behave as expected - within the error of measurement. Where the invariance principle is not instantiated in practice we would be motivated to examine the reason for that inadequate behavior and avoid using any such item in its current form. "This is the main reason that validity is an evolving property and validation is a continuing process." (Messick, 1995, p.741)

It would not be difficult to conclude that many users of the Rasch model are driven by the same sort of pragmatic considerations that guide model selection and statistic choice elsewhere in psychometrics. In his reply to my review (Bond, 2001) of his recent book, Michell (2001) asserted that such routine use of the Rasch model with data would not qualify as scientific measurement. My own first use of Rasch analysis was aimed at detecting whether two forms of assessment of cognitive development did indeed locate performances at the same Piagetian *stage*, or whether the high (Spearman's $\rho$) correlation between those assessments indicated merely similar ordinal relationships in the two datasets. I am now much more convinced that measurement in the human sciences must be theoretically driven. In common with the other quantitative rational sciences, we need theories of measurement of human variables which satisfy the requirements for scientific measurement. On the other hand, we need substantive theories about the human condition that allow us to examine how the responses that candidates make to our data collection devices are connected with the human attribute under investigation. While I might have been fortunate to have Piaget's 60 books and 600 articles on one side, and advice to use the Rasch model to solve developmental measurement problems on the other, many attempts at scale-building across the human sciences have a distinctly bottom-up approach.

My co-author, Christine Fox, characterizes part of her research background as that of an applied generalist. Much of her early work was spent providing psychometrics advice to statisticians and researchers in disciplines such as health education, counseling psychology, higher education, and early childhood, helping them to choose, analyze, and interpret a variety of statistical analyses. (Bond & Fox, 2001, p.xv) Her growing concern for the number of *post hoc* analyses she was obliged to conduct on data derived from poorly conceptualized questionnaires prompted her to adopt a new approach with faculty and students. She decided to ask them to explicate *a priori* the expected difficulty relationships among the items themselves, and between those items and the sorts of respondents for whom the data collection device was devised. The pre-requisite requirement was: Are all these of the same kind? *i.e.* are item and person performances indicative of the same underlying latent trait? This involved the ordering of key items from 'requiring just a little of this attribute', through 'requiring more of this attribute', to 'requiring a lot of this attribute'. Then her consultancy required colleagues to identify hypothetical respondents with various levels of the attribute under investigation, and to interpolate those persons into the item ordering (with a basic Guttman - style pattern in mind).

## The theory - practice dialectic

Leading Rasch measurement proponent, Ben Wright, claimed that the relative ease with which my students and I were able to develop sound scales of cognitive development at each attempt (*e.g.* Bond, 1995; Bond & Bunting, 1995; Bond 2001) was due largely to the wealth of theoretical underpinning of the 60 years of the Piagetian *oeuvre* from which our scale-building attempts derived. Fox's efforts with her colleagues are designed to prompt them to make explicit the implicit theories that they have about the way the human variables that concerns them are constructed, and revealed, in a variety of human test performances; a beginning attempt at the marriage of the theory of human measurement with a mini-theory about one human variable. This is a far more circumspect claim than the one routinely made in psychometrics. ". . . [P]sychologists have tended to overstate their achievements and to lose sight of where they are as a quantitative science." (Michell, 2003, p.299)

Such nascent attempts at theory building are designed to take account of common sources of invalidity following Messick: construct under representation and construct irrelevant variance are viewed as the two major threats to construct validity. First, Fox's advice provokes the colleague to consider the full extent of the variable, from its 'barely detectable' minimum to its 'off-the-scale' maximum. Second, the ordering prediction helps the investigator to reflect on the Rasch output in terms of the *a priori* statement of ordinal relationships; gaps in that correspondence might reveal levels of facility / difficulty that are unexpected in terms of the construct alone: *prima facie* candidates for construct-irrelevant-difficulty and construct-irrelevant-easiness. (Messick, 1995, p.742) Indeed, the Rasch /Messick link has been quite a rich source of ideas for those interested in the validity / measurement nexus (*e.g.* Fisher, 1994; Wilson, 1994). More recently, Smith (2001) discussed a raft of indicators of reliability and internal consistency from Rasch analysis and linked them quite directly to various aspects of Messick's unified validity framework.

## Rasch measurement theory in practice

If we were to take the example of the cognitive developmental test (BLOT) used to outline Rasch measurement principles with dichotomous data in Chapter 4 of Bond & Fox (2001), we might realize more readily how aspects of Rasch measurement might inform those crucial validity issues.

Figure 1 (from Bond & Fox, 2001, p. 40) provides a convenient summary of item statistics for the BLOT using a conception of the 'map' format which is original with that publication. On the vertical axis, the figure locates items according to item difficulty, while the size of the item marker indicates the relative precision (standard error) of that estimate. On the horizontal axis, the fit of those items is estimated in a standardized form as a $t$ distribution. Easier items are located at the bottom of the pathway, with more difficult items at the top. Items which misfit the model are located to the right; items which are closer to the deterministic, Guttman model are located to the left. While the 35 BLOT items are spread across the difficulty scale of about five logits, 30 of those items are compressed into a mere two logit space around the mean of the item difficulty scale.

This item distribution, of course, has implications for the sorts of decisions we

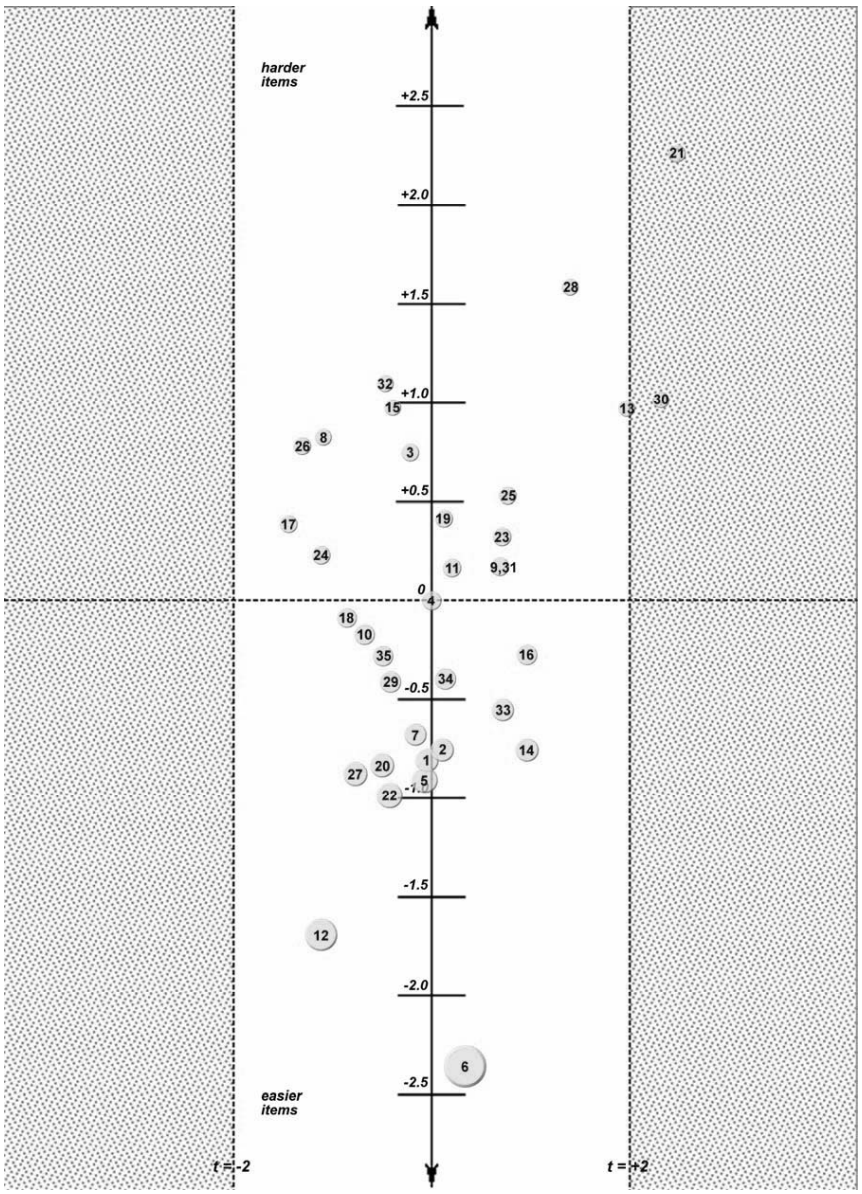


*Figure 1. BLOT item difficulties (Bond & Fox, 2001)*

are entitled to make about the students who take this test. Messick implicates this as one of the consequential aspects of validity. Decisions about students located at the very high and very low ends of the BLOT scale would, necessarily, be based on very small amounts of information derived from the three higher or two lower items. Decisions made about students at the midpoint of the scale, where the BLOT test is used to discern whether students are yet cognitively developed enough to benefit from the less concrete teaching typical in secondary and tertiary institutions, should be able to be made with a higher degree of confidence.

The fit indicators for the BLOT test revealed that two, perhaps three, items behaved more erratically than expected. The *prima facie* evidence suggests that counts for items 21 and 28 should not be routinely included in student BLOT *measures* without further reflection. Rasch measurement requires us to distinguish between counts (routinely summed raw scores on a test) and measures (linear interval measures based on the appropriate use of Rasch analysis) (Smith, 2001).

It almost goes without saying that the key Genevan text on the development of formal operational thinking (Inhelder & Piaget, 1958) contains very little information directly interpretable by a psychometrician about the relative difficulties of the logical operations identified as crucial to mature adolescent thought. Strikingly, those few indications that were made are nicely corroborated on the item difficulty map by the correspondences between logical pairs of items (Inhelder & Piaget, 1958, pp.293ff.) and Piaget's claims about excluding and testing false implications (Inhelder & Piaget, 1958, p.75) (see Bond, 1989). This sort of theory / practice evidence goes to the very heart of the issue of construct validity.

The more routinely used version of the variable map or Wright map from Rasch measurement (Figure 2), has the distinct benefit of aligning person performances as well as item performances on the same interval scale. The item-person map summarizing the performance of 150 English secondary school children on the BLOT test (Bond & Fox, 2001, p.42) confirms that many of these 14 year-olds are already beyond the upper limit of the test. In the BLOT test's usual role as a screening device to identify concrete operational as opposed to a formal operational thinkers, this is of little direct consequence.

However, the use of the BLOT test to measure changes in cognitive development over time (*e.g.* Endler & Bond, 2001) would be hampered significantly by this obvious ceiling effect. It reasonably might be contended that *construct under representation* would be unlikely in the BLOT; the item specifications were drawn from each and every descriptor found in the source text. The Rasch measurement evidence suggests, however, that the extent or range of formal operational thinking ability is under-represented. The most cogent critique of this shortcoming of the BLOT comes from my colleague Richard Meinhard who argues that Piaget was a 'systems theorist'. Meinhard always looks for evidence of the appropriate cognitive *system* at work in children's problem solving, and he contends that the mere sum of the individual system components is only a small part of the evidence of an interactive system at work.

So, at the more general level, Rasch measurement informs the quest for validity in future development of the BLOT test by pointing to the area requiring further items. At the more particular level, it behoves us to take the Rasch item difficulty information into consideration when item distractors are constructed. At present,

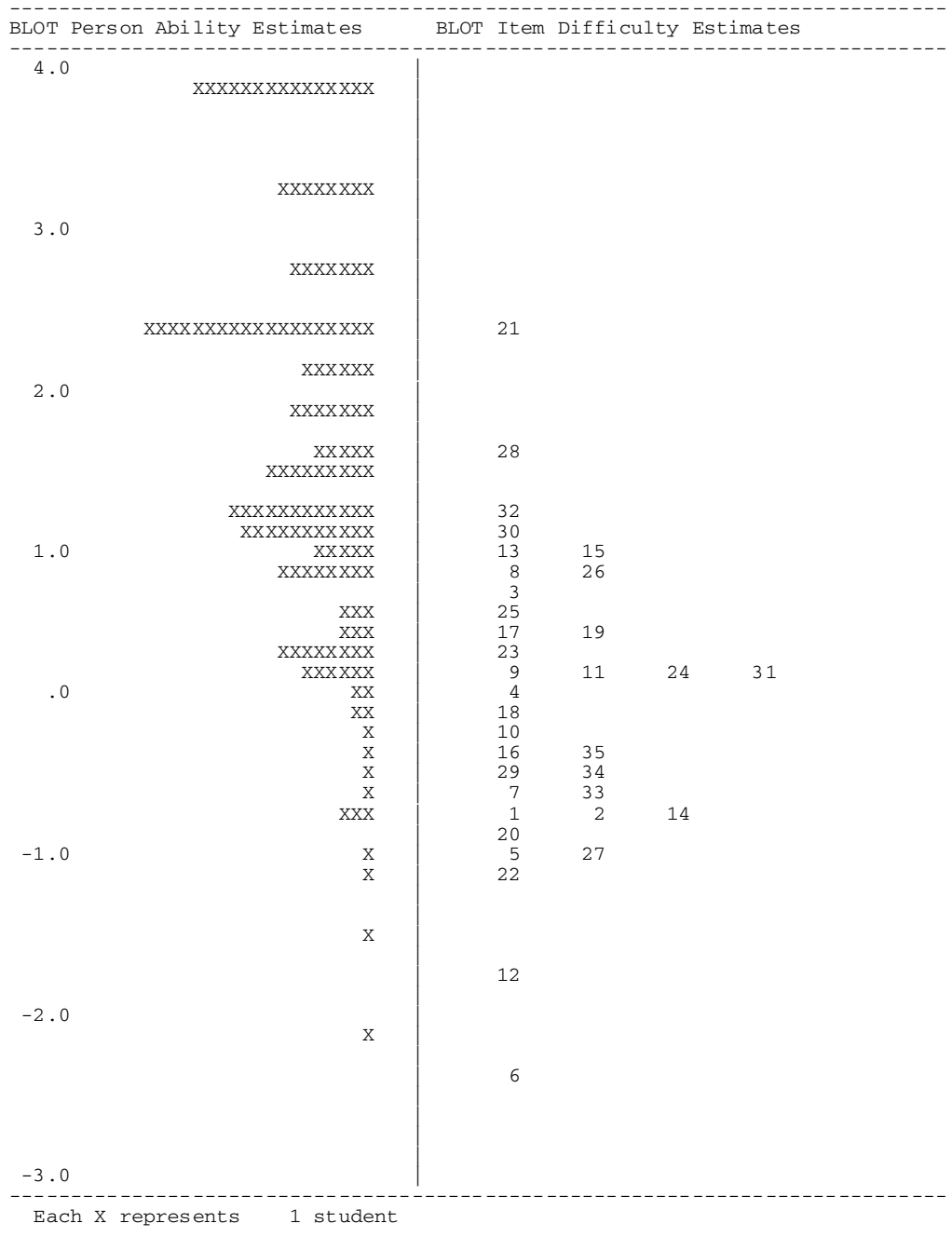


*Figure 2. 2 Item-person (Wright) map - BLOT*

the item distractors are other (incorrect) logical operations. In the future, the set of distractors for any one item could be constructed with an eye to their difficulty relative to that of the correct operational solution.

Rasch measurement indicators of item order and item fit then provide direct evidence about of the validity of the testing procedure, especially when the test content is explicitly embedded in substantive theory concerning the construct under investigation. And person ability and person fit indicators should be the subject of complementary considerations about the validity of person performance expectations derived from the construct.

## Person and Item Invariance

While we can easily see that the group of persons who took the test are a *sample* of the population of all possible test candidates, we less readily realize that the test items themselves are merely a sample of all possible test items. This is due, in no small part, to the relative ease of capturing a new sample of suitable test candidates and the relatively difficulty of constructing a new sample of suitable test items. In Rasch measurement, the principles and logic of analysis and interpretation for persons completely mirrors that for items. We have a constant reminder of the sample / population relationship for both items and persons, even though any sample's representativeness of the population is far less crucial.

The feature of parameter separation (Bond & Fox, 2001, pp.202-203; Smith, 2001) focuses on the property of the Rasch model that supports direct comparisons of person ability and item difficulty estimates, i.e. independently of the distribution of those abilities and difficulties in the particular samples of persons and items under examination. Ben Wright's challenge to those claiming to have a good test was forthright: divide your sample of persons in two according to ability and conduct item estimations for each half of the sample in turn; the relative difficulties of the items should remain stable across the two substantially different sub-samples. The corollary is also true: relative person ability estimates should remain invariant regardless of which half of the sample test items is used for the person estimation.

Figure 3 then, represents the results of my finally taking up Wright's challenge - rather than merely mouthing it and passing it on to others. I took the data from the 150 students who took the BLOT and divided them into two sub-samples: those with scores of 27-35 and the rest (*i.e.* raw scores 5-26). Each subsample was analysed separately and the 35 item estimates (and SEs) for each group were imported into an Excel spreadsheet (Bond & Fox, 2001, pp.62-65). The plot of item values show that, with the notable exception of item #21, the item estimates of the BLOT items are invariant (within error); the dotted line is not a regression line, but the Rasch modelled relationship required for invariance (the 95% control lines are based on the SEs for each of the item pairs).

Now, item 21 is, by far, the most difficult BLOT item (Figs 1 & 2) and while the high ability subsample allows us to gauge the difficulty difference between items 21 and 28 quite well, it is out of the range of the BLOT abilities shown by the low ability half of the sample. This results in the 21 - 28 interval being underestimated with the low ability group. Given that 26/35 has been estimated as the bound between concrete operational and formal operational performance on the BLOT, the invariance

of the item estimates (within error) helps affirm the integrity of the BLOT under Rasch procedures and demonstrates that the BLOT Rasch interval-scale maintains its properties across quite disparate sub-samples.

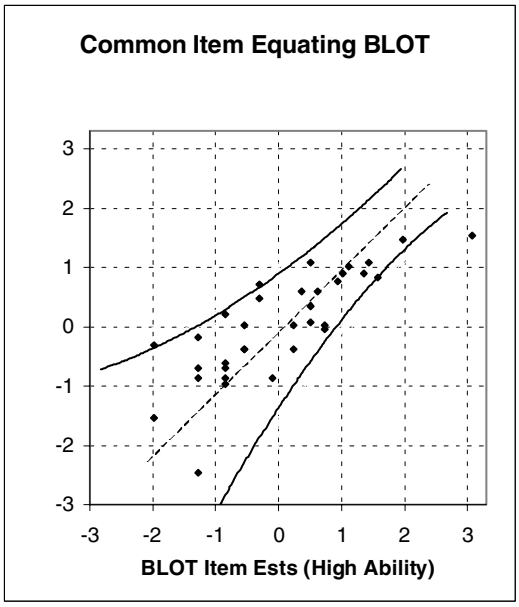


*Figure 3. Item Difficulty Invariance - Bond's Logical Operations Test*

## Towards equated universal measures

Along with a case built on face validity, concurrent validity has often been regarded as the *sine qua* non of validity evidence. Under True Score Theory, this evidence has usually been provided in terms of correlational evidence: a calculation of Pearson's $r$ or some form of factor analysis. In my early research years I often wondered about the reason for producing highly correlated tests. Unless the new test had some distinct benefit over the accepted standard in the field (*e.g.* faster, cheaper, ease of administration or marking), it seemed to me that the aim should be to produce a new test that was moderately related (prestige by association) but not so highly correlated that it could be regarded as redundant. But then, I could hardly have imagined hundreds of candidates sitting in front of computer screens all taking different versions of the same test: Computer Adaptive Testing, where the items chosen are dependent on the candidate's responses and the estimate of ability is required to be invariant across the particular sample of items delivered.

Moreover, we should expect that when one latent trait or underlying construct is the subject of a variety of tests, person measures should remain invariant (within error) across those testing conditions. The long-term aim of a genuine measurement system in the human sciences should be access to a series of co-calibrated testing situations, such that the raw score or locally calculated person measure could be expressed on the single scale used world-wide. A reasonable analogy from the physical sciences might be drawn from thermometry. Thermometers are used

to estimate temperature *indirectly*, by observation of the effect on some testing variable: the length of an enclosed column of coloured alcohol, the resistance of a metal wire, the straightening of a bi-metallic strip *etc.* The units on temperature are not additive in the physical sense (as are weight and length), and almost as many types of thermometer exist as there are temperatures to estimate. No one thermometer is useful for more than a few selective applications, and none measures over the full range of the temperature variable. Certainly, none does so without error. Even notwithstanding one major country's reluctance to use the international units of temperature measurement, translations from °F to °C are routine and - with the 'buyer beware' credo, "You get what you pay for!" in mind - temperatures are internationally measured, understood and compared after the most elementary sort of training in 'test administration'.

One object of Fisher's research (Fisher, 2000) is the establishment of what he terms 'universal metrics' for important aspects of human capital. His early attempts at equating rehabilitation scales (*e.g.* Fisher, Eubanks & Marier, 1997) revealed the remarkable utility of the Rasch model in supporting the co-calibration of instruments across samples, in spite of differences in number of items, as well as in the number and labels for Likert-scale options. Tesio and colleagues collaborated across national boundaries (between Italy and the USA) and equated Rasch rehabilitation measures on the FIM$^{TM}$ to reveal that variations in a few difficulty locations reflected differences between an incentive-driven medical system in the US and a more family-based support system in Italy (Tesio, Granger, Perucca, Franchignoni, Battaglia & Russell, 2002). Smith (2001) summarised the basic steps for co-calibration of test instruments while the key aspect of the conceptual basis for these procedures presented in Bond & Fox (2001), relies on earlier expositions (Wright & Stone, 1979; Masters, 1985; Masters & Beswick, 1986).

Common-person equating of two instruments involves a basic test of the invariance of person ability estimates of a single sample across the two testing situations. The two ability estimates for each person are plotted against each other; Rasch measurement prescribes that the plots should fall on a single line (with allowance made for the modelled SE pairs for each person locations). The converse case of common-test equating assesses the invariance of item estimates across two samples given the same test. Given that the Rasch model is quite robust in the face of missing data, comparative estimates can be made even for those persons who did not respond to all items, as well as for items attempted by only some of the respondents. (Chapter 5 of Bond & Fox (2001) provides details of a simple Excel spreadsheet to plot the estimates and associated Standard Errors.) To the extent that the Rasch model's requirements for measurement are satisfied, the relative abilities of two samples of persons on the two tests can be compared directly. When estimates on two remarkably different tests of formal operational ability were compared (for the N=150 sample above), the results showed the two tests to measure the same dimension, but that the PRTIII was much more difficult for high school students than was the BLOT (Bond, 1995).

The principle underlying the plots of person and item invariance across testing situations is exactly that underlying the detection of Differential Item Functioning (DIF). When an item's difficulty estimate location varies across samples by more than the modelled error, then *prima facie* evidence of DIF exists. Indeed, Scheuneman & Subhiyah (1998) from the National Board of Medical Examiners used merely an item

estimate difference greater than 0.5 logits as the criterion for detecting DIF in a 250 item medical certification test given to over 400 candidates. Given that SE estimates are inversely proportional to sample size, we could safely expect that a difference of 0.5 logit might have both statistical and substantive meaning (i.e. a genuine measured difference) on this high-stakes test. Their very basic approach, based on the violation of the Rasch model's expectation of estimate invariance, detected about 80% of the items uncovered by the Mantel-Haenszel procedure. Moreover, they argued, when the examiner understands the substantive construct under examination, and is thereby able to posit reasonable hypotheses about the performances of sub-samples of examinees, Rasch based indications of DIF are more directly comprehensible in terms of those group differences.

When we recall that the procedures for item and person estimation are mirrored, we must then also countenance the complementary situation of Differential Person Functioning: the examinee who performs differentially on two tests of the latent trait delivered in different contexts. Part of Bunting's research not reported in Bond & Bunting (1995) involved the equating of students' results derived from the individually delivered Piagetian interview of the pendulum task (Bond & Fox, 2001, Chapter 7) with those from the PRTIII - a demonstrated class-task of the same problem (Bond & Fox, 2001, Chapter 5). While the common person equating procedure did not disconfirm the presence of a single underlying construct, two sets of results were noteworthy, especially in terms of reflecting on the underlying developmental theory. First, the individual Piagetian interview version of the pendulum task was, on average about 2 logits easier for the sample than was the pencil and paper demonstrated class-task PRTIII - almost exactly the same difference and direction for the BLOT v. PRTIII comparison reported earlier (Bond, 1995; Bond & Fox, 2001, Chapter 5). Second, two of the DIF results highlight the complexities of measuring and understanding the complexities of the human condition. A male student, described by his teacher as coming from a male-dominated military family, scored much better on the PRTIII administered in the class setting by (older, male) Bond than he did in the individual Piagetian interview administered by (younger, female) Bunting. A female student performed much better in the individual Piagetian interview with Bunting than she did in the class group test where she seemed more focussed on her young male peer sitting (very closely) beside her. And, what is more, the person fit statistics in all four instances were quite unremarkable (Bunting, 1993). Now, reflect on that little theory / practice in context dialogue.

And, of course, while the many-facets Rasch model can be used to measure the systematic severity / leniency of judges (raters), Differential Rater Functioning would alert us to the presence of rater bias - the rater who (un)wittingly changes severity according some group characteristic. How is the theory of construct / context / rater going to deal with that sort of evidence?

## Objective Standard Setting

If, at this point, the Rasch model sets the standard for measurement in the human sciences, then surely the methods for setting pass marks or cut-scores for qualification examinations should reflect similar use of scientific approaches. Glass's scathing summary of the state of the cut-off or standard-setting art quarter of a

century ago seems to have garnered very little that could be regarded as progress since then. Glass argued that criterion-referenced standard setting approaches were merely a "psuedoquantification, a meaningless application of numbers to a question not prepared for quantitative analysis" (Glass, 1978, p.238). This sounds remarkably reminiscent of the scientific measurement lobby's critique of psychometrics in general (see Bond & Fox 2001; Michell, 1999). While the modified Angoff method seems to be the front-runner from the position of True Score Theory and IRT, the chief current proponent of the Objective Standards Setting approach both ably argues and successfully demonstrates the superiority of the latter approach over the former (Stone, 2001). Indeed, when the validity of standard setting is the key required outcome, then the OSS seems to have all the bases covered. The first steps towards this Rasch-based procedure were outlined in a public forum by Wright & Grosse (1993). Now, the cut-score is calculated by a series of logical steps developed by Stone for a series of consultancies to examining boards (Stone, 2002) and professional expositions (Stone, 2001), to include those abilities (items) deemed essential by the judges.

1 Judges use their expert judgement qualitatively to select their 'essential to qualify' items from all those on a Rasch-calibrated test.

2 The summary of each judge's selected items is quantified by calculating the mean item difficulty of those selected items.

3 The level of mastery required by the judges is expressed as a probability (*e.g.* 50%; 80% *etc.*) and converted to a logit measure by the usual odds to logits transformation (*e.g.* 60% = +.41 logits *etc.*)

4 As each person estimate has an associated SE, the judges must consider what to do when the error band of the candidate overlaps the cut score (so far calculated). Increasing the cut score by an error margin *protects the public* against an inadequate practitioner, while decreasing that score by an error margin *protects the innocent examinee* against the error-prone nature of testing and evaluation.

Benefiting from the iterative nature of Rasch intervals, the final cut-point for Objective Standard Setting is expressed as the sum of its component parts:
(Quantified Selected Content + Mastery Level) ± Confidence Level = Final Standard

The most convincing aspect of the OSS case as put by Stone (2001), is the comparison he made of passing rates on more than a dozen high-stakes certification examinations over three years: the stability of the OSS pass-rates was in sharp contrast to the relative instability under the modified Angoff procedure.

Objective Standard Setting then capitalizes on the two key attributes of a scientific measurement system in the human sciences: the validity of the test being used and the Rasch measurement properties of the resultant scale. And the consequence should be quite clear: unambiguous quantification of the qualitative evaluations made by certifying or standards setting boards about the competencies displayed by examinees on the qualifying exam. Setting a qualifying standard by adopting the OSS principles outlined above (and in more detail in Stone, 2001) provides a level of guarantee about examinee ability in relation to the standard not countenanced under the "70% is a pass" conclusion derived by any lesser standard setting technique.

## Conclusion

Messick reminded us that "validation is empirical evaluation of the meaning and consequences of *measurement*" (1995, p.747, emphasis added). This paper willingly accepts Michell's (1999) critique that psychometricians have overlooked (either wilfully or otherwise) the requirements of what measurement should look like in a rational quantitative human science. Indeed, Michell (1997; 2000) argued that psychology's persistent resistance to that critique seemed pathological; symptomatic of a disorder. Then it asserts that, in practical terms, the Rasch model provides practitioners with the most amenable practical guide to what scientific measurement should be like in the human sciences. The possible shortcoming of this approach in current practice is likely to be in the ability of misfit statistics to reveal where data from practice does not satisfy the measurement axioms derived from Luce *et al.* Moreover, it is argued that the invariance of item and person estimates across measurement contexts addresses the very essence of validity in the human sciences.

The Rasch measurement approach to the construction and monitoring of variable is so amenable to the issues raised in Messick's broader conception of construct validity that Smith (2001) draws direct one-to-one correspondences between the eight facets of construct validity from Messick (1995) and inferences drawn directly from the theory and practice of Rasch measurement. The theory / practice dialectic (Bond & Fox, 2001) ensures that validity is foremost in the mind of those developing measures and that genuine scientific measurement is foremost in the minds of those who seek valid outcomes from assessment. Failures of invariance, such as those referred to as DIF, alert us to the need to modify our assessment procedures or our substantive theory about the human condition, or both. In Objective Standard Setting, Stone (2001) demonstrated how valid pass-fail judgements reflect both on the conception of the underlying latent trait and the transparent measurement procedures by which experts evaluations are transformed to performance standards.

Rasch measurement instantiates an approach to assessment which can be described, borrowing Messick's own words, as a "comprehensive view of validity [which] integrates considerations of content, criteria, and consequences into a construct framework for empirically testing rational hypotheses about score meaning and utility." (Messick, 1995, p.742)

## References

Bond, T. G. (1989) An investigation of the scaling of Piagetian formal operations. In P. Adey (Ed.) *Adolescent Development and School Science.* London: Falmer.

Bond, T. G. (1995). *BLOT: Bond's logical operations test.* Townsville: James Cook University.

Bond, T.G. (2001) Book Review 'Measurement in psychology: A critical history of a methodological concept'. *Journal of Applied Measurement, 2, 1,* 96-100.

Bond, T. G., & Bunting, E. M. (1995). Piaget and measurement III: Reassessing the méthode clinique. *Archives de Psychologie, 63(247),* 231-55.

Bond, T.G. & Fox, C. M. (2001) *Applying the Rasch model: Fundamental measurement in the human sciences.* Mahwah, N.J.: Erlbaum.

Bunting, E. (1993). *A qualitative and quantitative analysis of Piaget's control of variables scheme.* Unpublished thesis, Townsville: James Cook University.

Endler, L.C. & Bond, T.G. (2001).Cognitive development in a secondary science setting. *Research in Science Education, 30, 4,* 403-416.

Fisher, W. P., Jr. (1994). The Rasch debate: Validity and revolution in educational measurement. In M. Wilson (Ed.), *Objective measurement: Theory into practice.* Vol. II (pp. 36-72). Norwood, New Jersey: Ablex Publishing Corporation.

Fisher, W. P., Jr. (2000). Objectivity in psychosocial measurement: What, why, how. *Journal of Outcome Measurement, 4, 2,* 527-563.

Fisher, W. P., Jr., Eubanks, R. L., & Marier, R. L. (1997). Equating the MOS SF36 and the LSU HSI physical functioning scales. *Journal of Outcome Measurement, 1, 4,* 329-362.

Glass. G.V. (1978). Standards and criteria. *Journal of Educational Measurement, 15,* 237-261.

Inhelder, B., & Piaget, J. (1958). *The growth of logical thinking from childhood to adolescence.* London: Routledge & Kegan Paul. (Original work published in 1955.).

Karabatsos, G. (1999a). *Rasch vs. two- and three-parameter logistic models from the perspective of conjoint measurement theory.* Paper presented at the Annual Meeting of the American Education Research Association. Montreal, Canada.

Karabatsos, G. (1999b). *Axiomatic measurement theory as a basis for model selection in item-response theory.* Paper presented at the 32th Annual Conference for the Society for Mathematical Psychology. Santa Cruz, CA.

Karabatsos, G. (2000). A critique of Rasch residual fit statistics. *Journal of Applied Measurement, 1, 2,* 152-176.

Luce, R.D., & Tukey, J.W. (1964). Simultaneous conjoint measurement: A new type of fundamental measurement. *Journal of Mathematical Psychology, 1, 1,* 1-27.

Masters, G. N. (1985, March). Common-person equating with the Rasch model. *Applied Psychological Measurement, 9,* 73-82.

Masters, G. N., & Beswick, D.G. (1986). *The construction of tertiary entrance scores: Principles and issues.* Melbourne: Centre for the Study of Higher Education, University of Melbourne.

Messick, S. (1995) Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist, 50, 9,* 741 - 749.

Michell, J. (1997). Quantitative science and the definition of measurement in psychology. *British Journal of Psychology, 88, 3,* 355 - 383.

Michell, J. (1999). *Measurement in psychology: A critical history of a methodological concept.* Cambridge: Cambridge University Press.

Michell, J. (2000). Normal science, pathological science and psychometrics. *Theory and Psychology, 10,* 639-667.

Michell, J. (2001). Response to Trevor Bond's book review: Rasch models and the prospect of psychological measurement. *Journal of Applied Measurement, 2, 2,* 201-204.

Michell, J. (2003). Measurement: A beginner's guide. *Journal of Applied Measurement, 4, 4,* 298-308.

Rasch, G. (1960). Probabilistic models for some intelligence and attainment tests. Copenhagen: Danish Institute for Educational Research. (expanded edition, 1980, Chicago: The University of Chicago Press).

Scheuneman, J.D. & Subhiyah, R.G. (1998). Evidence for the validity of a Rasch model technique for identifying differential item functioning. *Journal of Outcome Measurement, 2, 1,* 33-42.

Smith, Jr., E.V. (2001). Evidence for the reliability of measures and validity of measure interpretation: A Rasch measurement perspective. *Journal of Applied Measurement, 2,* 281-311.

Stevens, S. S. (1946). On the theory of scales of measurement. *Science, 103,* 667-680.

Stone, G. (2001). Objective standard setting (or truth in advertising). *Journal of Applied Measurement, 2,* 187-201.

Stone, G. (2001). Objective standard setting. Paper presented at the annual meeting of the National Organization for Competency Assurance, New Orleans, LA, December 2001.

Stone, G. (2002). The Emperor has no clothes: What makes a criterion-referenced standard valid? Paper presented at the Fifth Annual International Objective Measurement Workshop, New Orleans, LA, April, 2002.

Tesio, L., Granger, C.V., Perucca, L., Franchignoni, F.P., Battaglia, M.A. & Russell, C.F. (2002). The FIMTM instrument in the United States and Italy: A comparative study. *American Journal of Physical Medicine & Rehabilitation, 81, 3,* 168-176.

Wilson, M. (1994). Comparing attitude across different cultures: Two quantitative approaches to construct validity. In M. Wilson (Ed.), *Objective measurement: Theory into practice,* Volume 2 (pp. 271-294). Norwood, New Jersey: Ablex.

Wright, B.D. and Grosse, M. (1993). How to set standards. *Rasch Measurement Transactions* of the Rasch Measurement SIG, American Educational Research Association, 7, 3, 315-316.

Wright, B.D. and Stone, M. H. (1993). *Best test design.* Chicago: MESA Press.