

What are the Effects of Allowing Crossing Item Characteristic Curves into our Measurement Model?

Timothy W. Pelton, University of Victoria, tpelton@uvic.ca

DRAFT

Ideal measurement requires unidimensionality, additivity and objectivity. If this ideal could be achieved in the human sciences, it is expected that observations of the ability of persons with respect to items would be consistent with a stochastic approximation to Guttman's scaling (Guttman 1944). Rasch presented a model that is built upon these expectations and described it with respect to probabilities:

A person having a greater ability than another should have the greater probability of solving any item of the type in question, and similarly, one item being more difficult than another one means that for any person the probability of solving the second item correctly is the greater one. (Rasch 1960, 1980, 1993 1993 #52) as cited in (Wright 1999).

Wright (1999) has nicely stated the assumptions, process and utility of Rasch measurement in the human sciences. As part of his exposition, Wright describes some of the 'mistakes' (p. 65) that scientists have been making in their attempts at measurement and outlines the requirements for success in measurement. Among these principles Wright includes (a) avoid the use of models that fail to converge such as the two-parameter logistic (2PL) and three-parameter logistic (3PL) item response theory (IRT) models; (b) use models that have sufficient statistics to allow for independent estimation of parameters; and (c) do not use models that "cause the hierarchy of relative item difficulty to change at every ability level" (Wright, 1999 p. 95). These principles are each related to the assumption of parallel item characteristic curves (ICCs) that is in turn the foundation of the Rasch model.

When the underlying structure of the data is consistent with non-crossing ICCs, conjoint additivity (Luce and Tukey 1964) is possible and when the data are consistent with parallel logistic ICCs the traditional Rasch model can claim to produce interval level estimates of person ability and item difficulty (Brogden 1977; Perline, Wright et al. 1979). Other justifications for seeking parallel ICCs in our measurement model include simplifying the meaning of estimates of ability relative to item difficulty as suggested by Wright (see principle c above). Depending on one's interpretation of the unidimensionality,

parallel ICC's may or may not be implied in the more commonly stated IRT requirements of unidimensionality, local independence and fit to the theorized item characteristic curve (ICC).

The principles given by Wright are requirements for the ideal application of the Rasch Model. When these principles are mirrored in the situation and the sample size is adequate, then the noise will be minimal and very stable and accurate measurement scales will result. A good model-reality fit is often found in mature physical science domains (e.g. Newtonian physics) where the models are much more stable and the unknown factors have generally been either identified and included or controlled. Although it is likely that there are remaining undetected factors or other forms of misfit, their effect on the system is minimal in the normally observed ranges. The success realized in measurement in the physical sciences has occurred because of the identification of a substantial foundation of "extensive" constructs (e.g., length, mass, time) that fit reality and because all "intensive" (e.g., density, temperature) (Cohen and Nagel 1934) constructs are effectively defined in relation to these extensive constructs.

The suitability of these principles to measurement situations with much higher levels of noise (i.e., in the human sciences) is less certain. Item sets (or observation sets) will always be affected by varying levels of noise resulting from the presence of unrecognized or uncontrollable secondary factors and item specific factors. In a unidimensional measurement model, secondary factors may also be described as local dependencies or as dependent multidimensionality – this type of multidimensionality is a common source of noise and is explored elsewhere (McDonald 1981; Hattie 1984; Hambleton and Rovinelli 1986). The item specific factors may also be described as uniqueness components or independent multidimensionality – this type of multidimensionality affects items independently resulting in varying discriminations and crossing ICCs. This independent multidimensionality is not permitted in the Rasch model, but it is embraced in the 2PL and 3PL models (McDonald 1999; Pelton and Bunderson 2002).

Lord (Lord 1980) and McDonald (McDonald 1999) have both suggested that practical measurement models need to accommodate uniqueness – or independent multidimensionality – when it is present in the items. The two-parameter logistic (2PL) model (Birnbaum 1968) accommodates this type of multidimensionality by allowing for variations in the ICC slopes that in turn result in crossing ICCs. With these crossing ICCs, the 2PL model cannot claim conjoint transitivity and thus cannot claim conjoint additivity and thus it does not have the potential to yield perfectly interval measurement scales. Yet

intuitive exploration of the model leads to an understanding that the constrained iterative estimation process will yield something close to interval measurement when the structures underlying the data are consistent with the model.

Evidence suggests that the relative accuracy of person ability estimates is remarkably consistent across models regardless of the model-reality misfit while the accuracy of item difficulty estimates appears to be highly dependent upon the measurement model used and the amount of misfit in the structure underlying the data set (Pelton and Bunderson 2002). Although it is obvious that the Rasch model is a measurement ideal, and will produce the most accurate results when presented with data derived from an item set with an underlying structure consistent with the model, it is not obvious that this same model will produce superior results when presented with less than ideal data from a reality that does not fit the model assumptions. Indeed, previous results suggest that the Rasch model advantage over the 2PL model with respect to the accuracy of item difficulty estimates diminishes very rapidly, and disappears, as the data deviates from the Rasch model assumptions of unidimensionality and no guessing effects (Pelton and Bunderson 2002). When the true underlying structure of a data set contains crossing ICCs, which model will maximize the accuracy of the estimates of the item positions as a quasi-interval measurement scale is constructed?

One approach to working with imperfect data (as would be generated by a misfitting reality) is to continue using the Rasch model. This approach requires the implicit hope that the Rasch model advantage with respect to its link to true interval scale estimation through conjoint additivity is sufficient to allow it to produce good results even when the underlying structure of the data is moderately inconsistent with the model assumptions. However, it is understood that when the underlying structure of the data set is not perfectly consistent with the Rasch assumptions (i.e., in every case in the human sciences), the results are, in fact, on a quasi-interval scale that can only approximate the true latent trait positions (Karabatsos 2001)

Alternatively the 2PL model might be used to estimate item and person measures. Because the 2PL model can accommodate the varying discriminations (independent multidimensionality), there will be a better model-reality fit. The lack of sufficient statistics for the 2PL model estimation process requires the use of arbitrary constraints in the calibration process to ensure convergence thereby introducing some unknown degree of error to the item difficulty and person ability estimates (Mislevy and Bock 1990;

Wright 1991). Regardless of the underlying structure defining the data (i.e., even if the underlying structure generating the data were consistent with the Rasch model requirements), the 2PL model can only claim to produce item difficulty estimates on a quasi-interval scale.

Efforts to refine an item set might yield some reduction in the effects of unwanted multidimensionality (dependent or independent), but when too much emphasis is placed on the selection of items that correlate closely with a common factor, the validity of the measure being created may be reduced (Bond and Fox 2001). Thus it may be reasonable to conclude that for valid measurement in the human sciences, independent and dependent multidimensionality will always be present to some degree and that the true underlying structure of any nontrivially distinct set of items or observations (e.g., a test) contains crossing ICCs.

Because crossing ICCs preclude the attainment of perfect interval measurement scales in the human sciences a need arises to assess the relative accuracies of the Rasch and 2PL measurement models as the degree of independent multidimensionality and the availability of information fluctuates.

In Figures 1 and 2, scatterplots are presented that contain the Rasch and 2PL calibrated item difficulty estimates from the 25 datasets that were replicated using the same underlying item parameters and an offset ability distribution corresponding to a very capable person sample. While the underlying parameters are more extreme than would typically be found in practice, Figure 1 effectively demonstrates that the Rasch model item estimates are person sample dependent when the underlying structure of the items from which the data set is determined is inconsistent with the model assumptions (i.e., a bias may be introduced).

In Figure 1 (Pelton 2002) the narrow distribution of item difficulty estimates across replications indicates that the Rasch model produces fairly reliable or stable results, but that the deviations from the true values are substantial (i.e., they are inaccurate) when items are subject to independent multidimensionality. The reported standard errors were found to be good approximations to the standard deviation of the estimates across replications but because the error estimates do not include the effects of misfit they are substantially smaller than the actual root mean square deviation (RMSD) from the generating difficulty. The errors increase in relation to the distance between the item difficulty and the mean of the person sample ability increases indicating that the error or bias is person sample dependent.

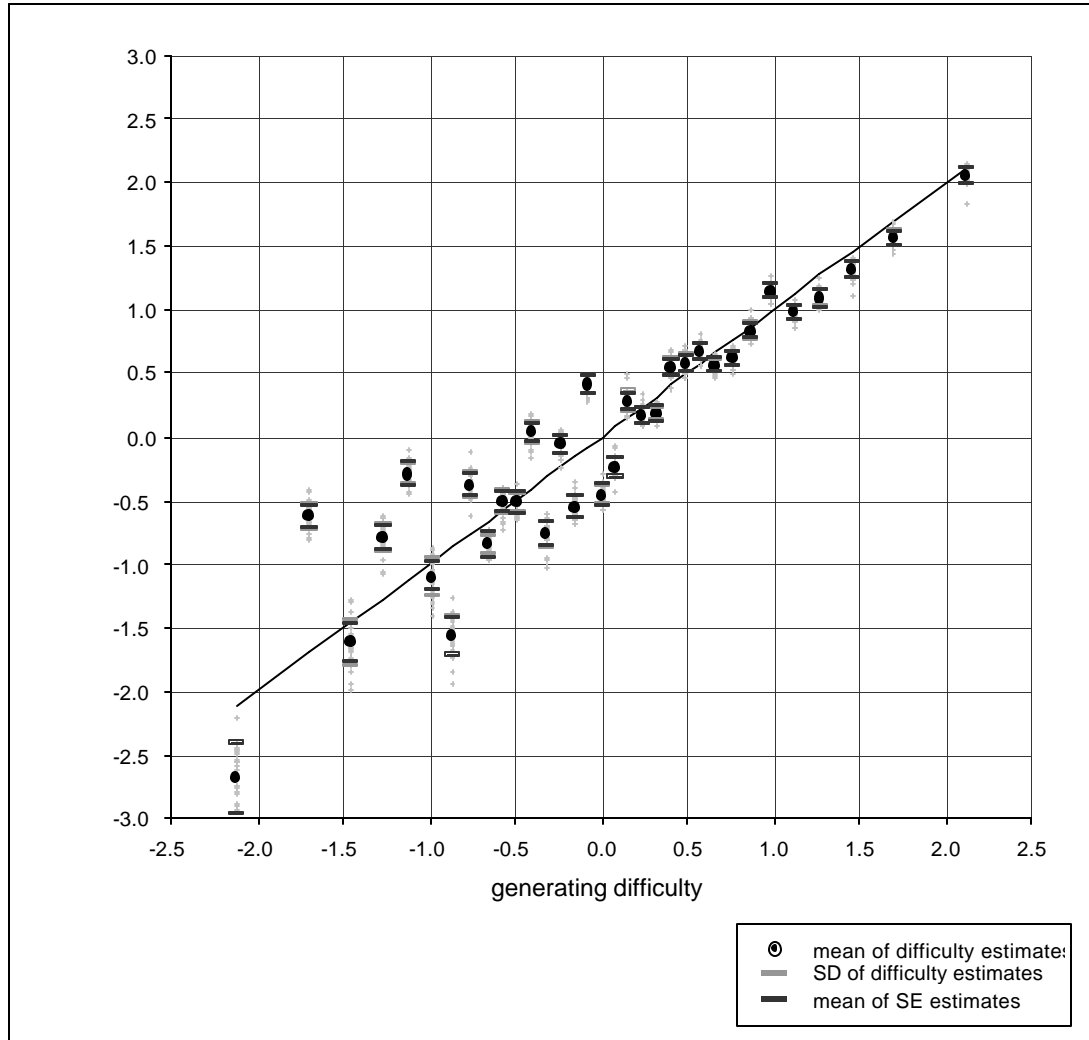


Figure 1. A comparison of Rasch estimates of item difficulty across 25 replications to the generating or true difficulty using an off-target ability sample ($\theta = N(1.5, 1)$), with an item structure containing moderately varying discriminations ($a = N(0.8, 0.3)$) and minimal pseudo-guessing ($c = N(0.05, 0.02)$) (Pelton 2002).

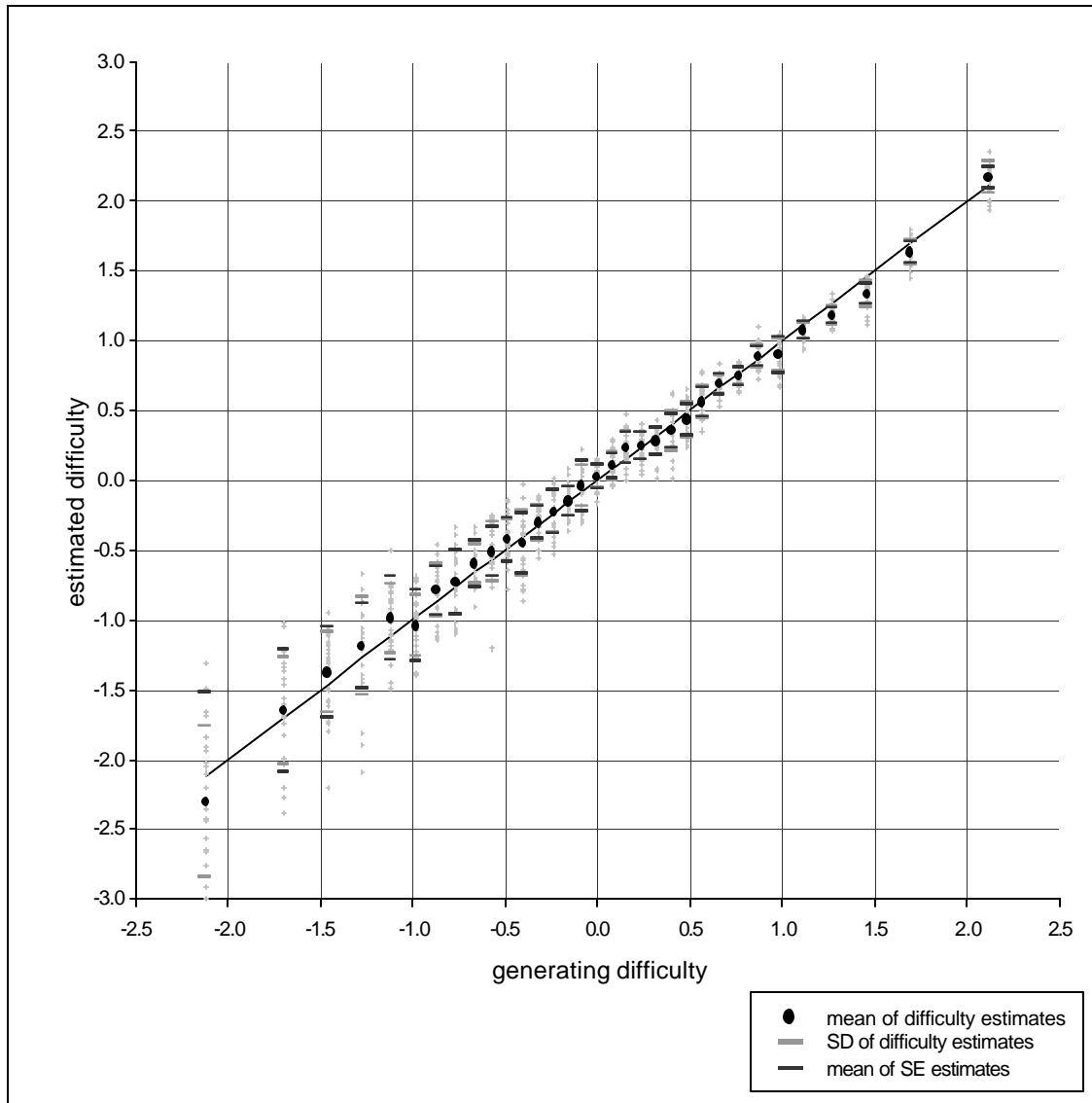


Figure 2. A comparison of 2PL estimates of item difficulty (across 25 replications) to the generating or true difficulty using an off-target ability sample ($\theta = N(1.5, 1)$), with moderately varying discriminations ($a = N(0.8, 0.3)$) and minimal pseudo-guessing ($c = N(0.05, 0.02)$) (Pelton 2002).

In Figure 2 (Pelton 2002), it can be observed that the 2PL model produces more accurate results for the items, although the stability of these estimates is diminished (an expected effect of estimating a second parameter). The error estimates from the 2PL model were much more consistent with the observed RMSD, although still somewhat smaller. Again the size of the error and the deviation of the mean difficulty estimate increased as the distance from the mean of the population distribution increased.

These results are at once consistent and inconsistent with Wright's suggestion that crossing ICCs would lead to reduced meaning (Wright, 1999). They are consistent in terms of anticipating the difficulties the Rasch model will have when the underlying structure of the item set used contains such varying discriminations. The results are inconsistent with the suggestion that crossing ICCs would lead to reduced meaning when the Rasch and the 2PL models are used to estimate item positions using the same data. It appears that the positions of items can be approximated more accurately – and thus more meaningfully – by the 2PL model because the model is able to accommodate independent multidimensionality.

Figure 3 shows a slightly modified conjecture derived from earlier results (Pelton 2002). The back plane of this conjecture (information x independent multidimensionality) suggests that in the absence of guessing, the Rasch model will produce superior results when either the sample is small or when the sample is large and the degree of independent multidimensionality is small, while the 2PL model will produce superior results when the sample is larger and the items are affected by higher degrees of independent multidimensionality.

This Monte Carlo study was designed to explore the relationship between the accuracy of Rasch and 2PL model item difficulty estimates while the amount of information and the degree of independent multidimensionality are varied (the back plane of the conjecture in Figure 3). To accomplish this, a series of simulations were conducted where the number of 'persons' in the sample (information) and the degree of independent multidimensionality (variance of discrimination parameters) was systematically varied. It is hoped that the results of this study might assist researchers in the design of their experiments (sample size and distribution) and in the selection of a measurement model that is most appropriate when their data sets contain misfit due to item uniqueness.

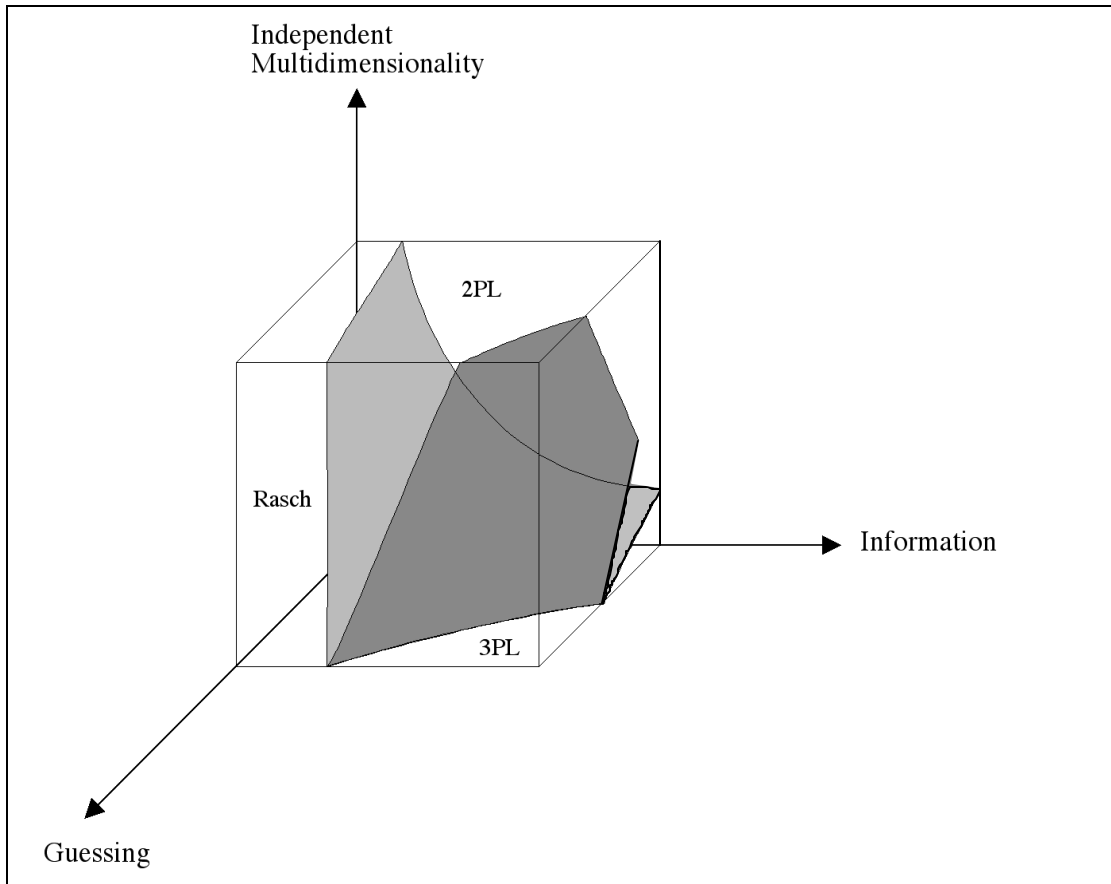


Figure 3. Conjectured regions of superior accuracy of item difficulty estimates for various Rasch and IRT models in the presence of varying degrees of information (sample size), independent multidimensionality (degree of variation of discrimination) and pseudo-guessing (Pelton 2002).

Method

The parameter distribution combinations chosen here are intended to allow for the systematic observation of effects of sample size (information) and discrimination distribution (independent multidimensionality). Each parameter combination distribution was selected to be within the realm of possibility in measurement applications.

Discrimination parameter distributions in this simulation study are set to constant, $U(0.8)$, representing parallel ICCs, mild, $N(0.8,0.1)$, moderate, $N(0.8,0.2)$, and high, $N(0.8,0.3)$ (on the probit scale). These distributions are fairly consistent with Linacre's conclusion that "...central discriminations from 0.5 to 1.7 (on the logit scale) produce good fit to the Rasch model..." (Linacre 2000) and with previously observed distributions of discriminations (e.g., (Hambleton, Swaminathan et al. 1991)).

Sample sizes include large-classroom size through large-scale calibration size ($N= 100, 500, 2500$ and 12500). Item sample size ($K=50$), item difficulty and person ability parameter distributions $N(0,1)$ are held constant in this simulation and guessing is not modeled ($c=0$). Each of these choices was intended to yield a data model that was moderately conservative and comparable with many empirical observations.

The simulation uses the 2PL model to estimate the probabilities of correct responses for each person-item pair creating a person x item probability matrix for each parameter distribution combination.

$$P(X_{ij} = 1 | \mathbf{q}_i, a_j, b_j) = \frac{1}{1 + e^{-Da_j(\mathbf{q}_i - b_j)}} . \quad (1)$$

Where:

\mathbf{q}_i is the ability of person i ,

a_j is the discrimination that reflects the degree of independent multidimensionality

b_j is the difficulty of item j , and

D is the a scaling factor to allow the approximation of a normal distribution,

Twenty-five response matrices were replicated from each response probability matrix by comparing random numbers to each cell in the probability matrix with the result being set to one (success) when the random number was less than the probability and zero (failure) otherwise. 400 data files were generated in all (4 sample sizes x 4 discrimination levels x 25 replications). Each of these data files was

processed twice – once with BIGSTEPS (Linacre and Wright 1999) to generate the Rasch estimates and once with Bilog (Mislevy and Bock 1990) to generate the 2PL estimates. Both programs were applied using default settings – with the exception that the sample size was explicitly set in Bilog in order to avoid the default sampling of 1000 persons when more persons were available, and fixing the number of EM cycles (20), Newton-Gauss iterations (10), and quadrature points (20)

RMSD across items was used to estimate the accuracy of standardized item difficulty estimates for each replication and each model using 100% of the items, the middle 80%, the middle 60%, the middle 40% and the middle 20% of the items. As an example, the $RMSD_{items_k, 60\%}$, which is an estimate of the error on replication k for the middle 60% of the items in the set was calculated using:

$$RMSD_{items_k, 60\%} = \sqrt{\frac{\sum_{j=11}^{40} (\hat{b}_{jk} - b_{jg})^2}{30}} \quad (2)$$

where: \hat{b}_{jk} is the difficulty estimate for item j in replication k , and
 b_{jg} is the generating difficulty for item j .

Scatter plots of item difficulty estimates were produced to allow for examination and comparison of results.

The mean RMSDs across replications were used to estimate points for error contours for both the Rasch and the 2PL models. The estimation of these contour points was based on linear interpolation on the standard deviation of the item discriminations and a square root transformation on the person sample size (after initial probing indicated such might be appropriate). The intersections of these two of error contour families (Rasch and 2PL) were plotted and a curve was estimated (using Excel's power trendline) to demark the regions of item difficulty accuracy estimate superiority for the two measurement models (Rasch and 2PL).

Finally, in an attempt to generate a heuristic to guide practitioners the Rasch results were further examined. Specifically, the mean standard deviations of the standardized Infit and Outfit statistics (Linacre 2001) across items and across replications from the BIGSTEPS (Rasch) estimates were examined in relation to directly observed model superiority and the curve demarking model superiority.

Results

Figure 4 presents the combined scatter-plots of the Rasch model item difficulty estimates compared to generating difficulty over 25 replications with a sample size of 100 and parallel ICCs ($a=U(0.8)$). Figure 5 presents the corresponding scatter-plot of item difficulty estimates using the 2PL model. As would be expected (because of the perfect fit of the generating reality to the measurement model), the Rasch model results (mean RMSD=0.19) are superior to the 2PL model results (mean RMSD=0.23) in this case.

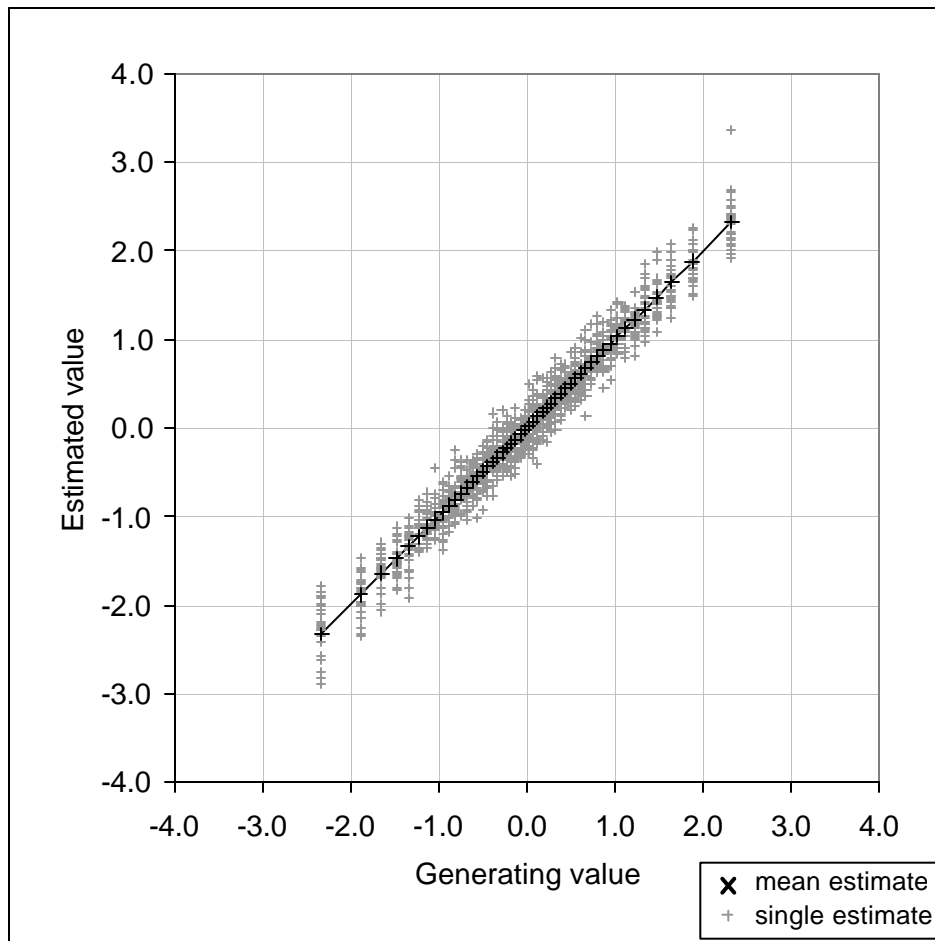


Figure 4. Comparing the standardized Rasch item difficulty estimates of 25 replications to the generating item difficulties with parallel ICCs and a small sample size ($a=U(0.8)$, $N=100$).

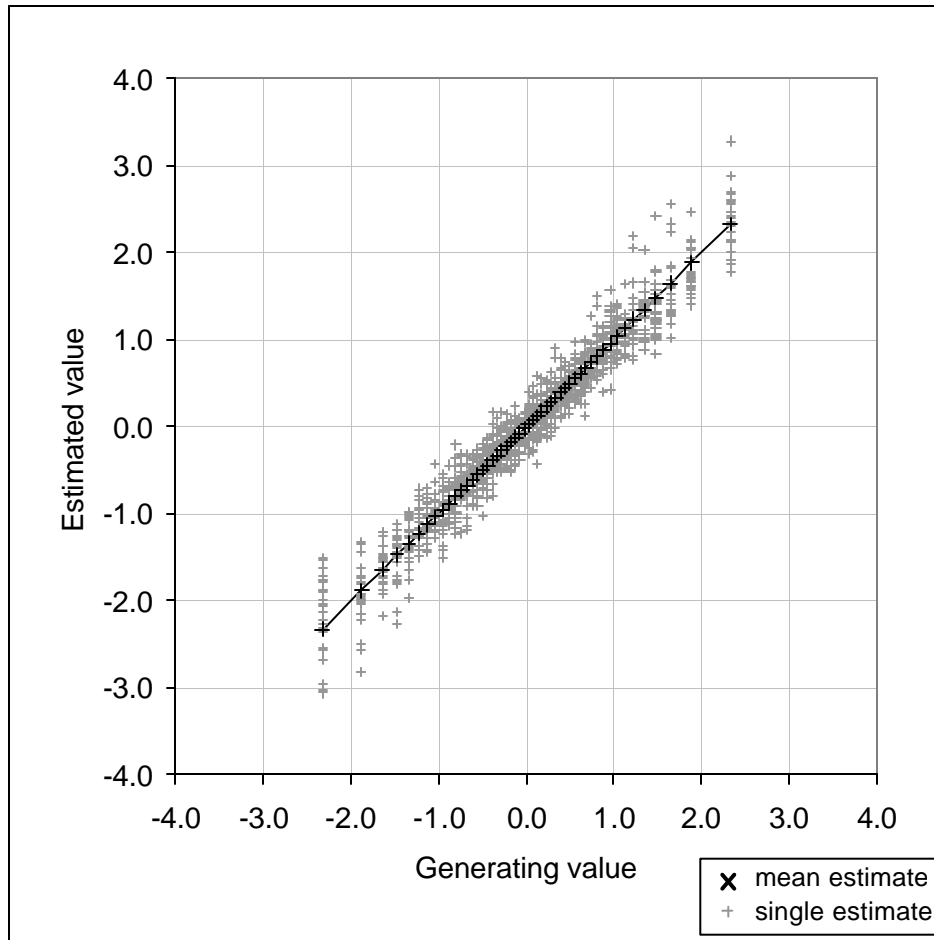


Figure 5. Comparing the standardized 2PL item difficulty estimates of 25 replications to the generating item difficulties with parallel ICCs and a small sample size ($a=U(0.8)$, $N=100$).

Figures 6 and 7 present the combined scatterplots for the Rasch and 2PL models where the sample size is 100 and the standard deviation of the discrimination is 0.3 ($a=N(0.8,0.3)$). Here, the Rasch model exhibits a tendency to report biased estimates of item difficulty when the item difficulty is off-target from the mean of the person ability distribution and the discrimination is substantially different from the mean discrimination. Although it is not visually obvious which of the two models should be preferred in this case, the 2PL model results (mean RMSD=0.25) were found to be superior to the Rasch model results (mean RMSD=0.31). It should be noted that Bilog (2PL) failed to generate error estimates on one of the 25 data sets – although the item difficulty estimates appeared reasonable in that instance and were included in our analysis.

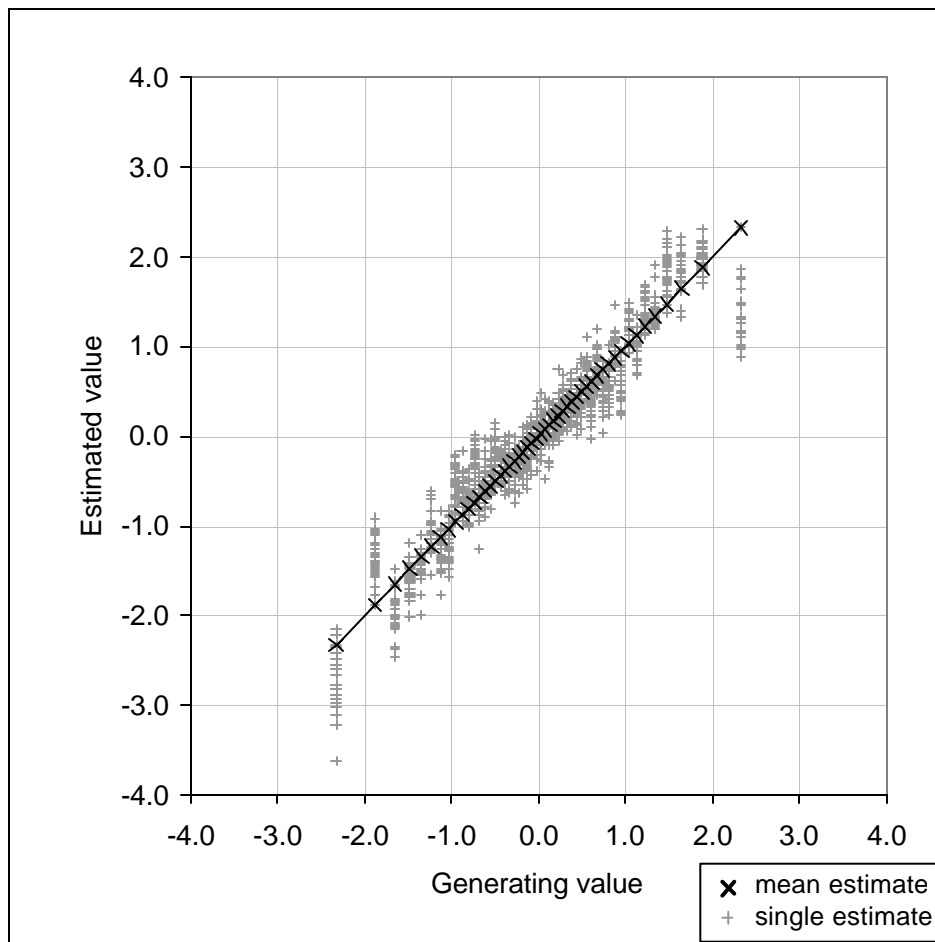


Figure 6. Comparing the standardized Rasch item difficulty estimates of 25 replications to the generating item difficulties with non-parallel ICCs and a small sample size ($a=N(0.8,0.3)$, $N=100$).

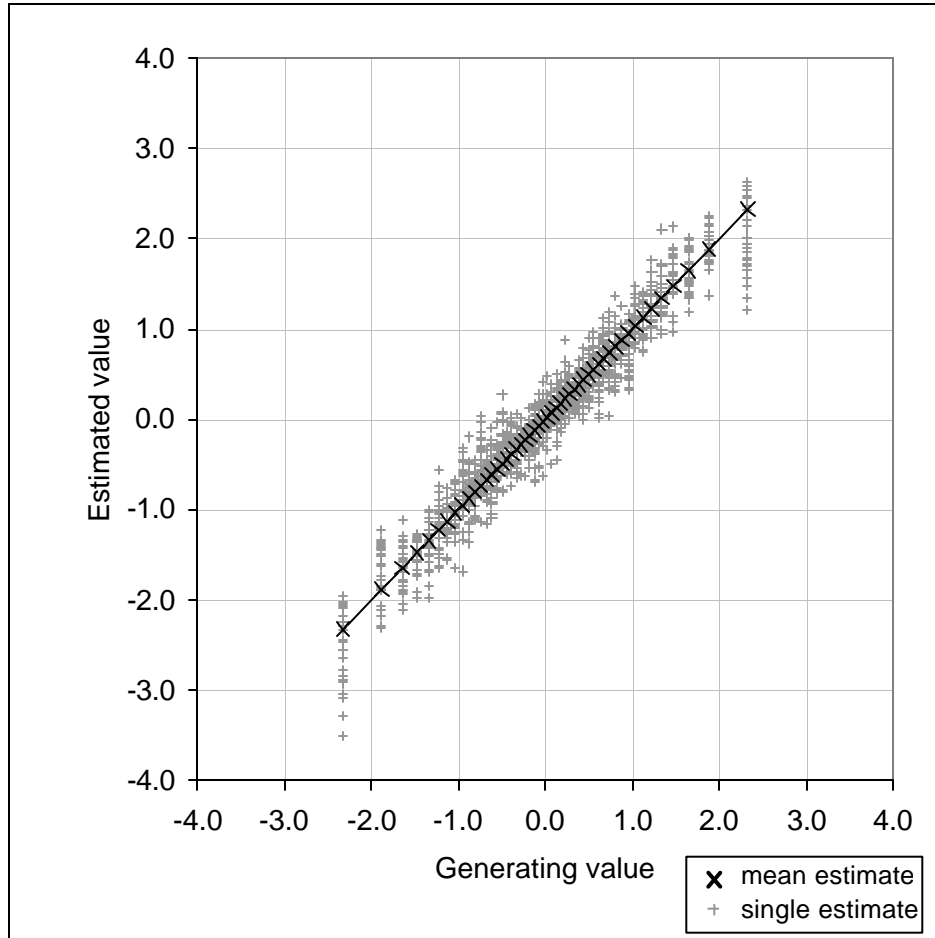


Figure 7. Comparing the standardized Rasch item difficulty estimates of 25 replications to the generating item difficulties with non-parallel ICCs and a small sample size ($a=N(0.8,0.3)$, $N=100$).

Figures 8 and 9 present the combined scatterplots for the Rasch and 2PL models where the sample size is 12,500 and the generating discrimination is constant ($a=U(0.8)$). Here, as should be expected, both models exhibit very accurate item difficulty estimates. The Rasch model, by virtue of its simpler estimation process produces more accurate item difficulty estimates (mean RMSD=0.017) than the 2PL model (mean RMSD=0.026).

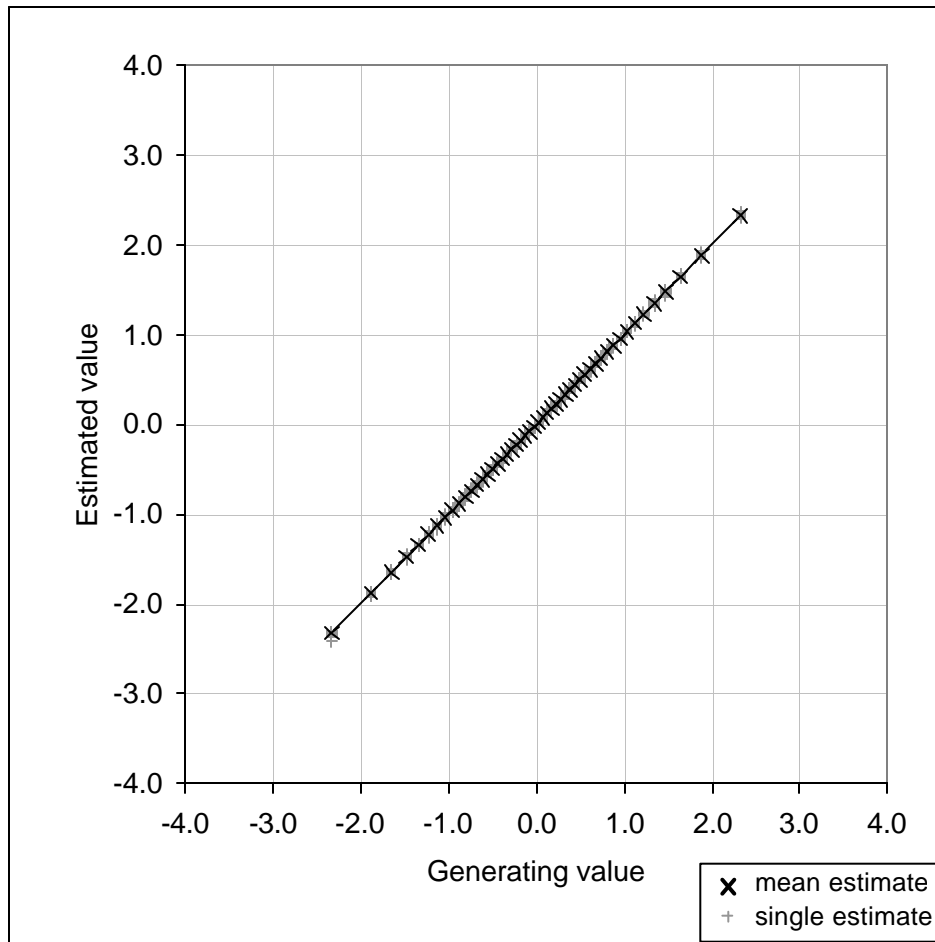


Figure 8. Comparing the standardized Rasch item difficulty estimates of 25 replications to the generating item difficulties with parallel ICCs and a large sample size ($a=U(0.8)$, $N=12500$).

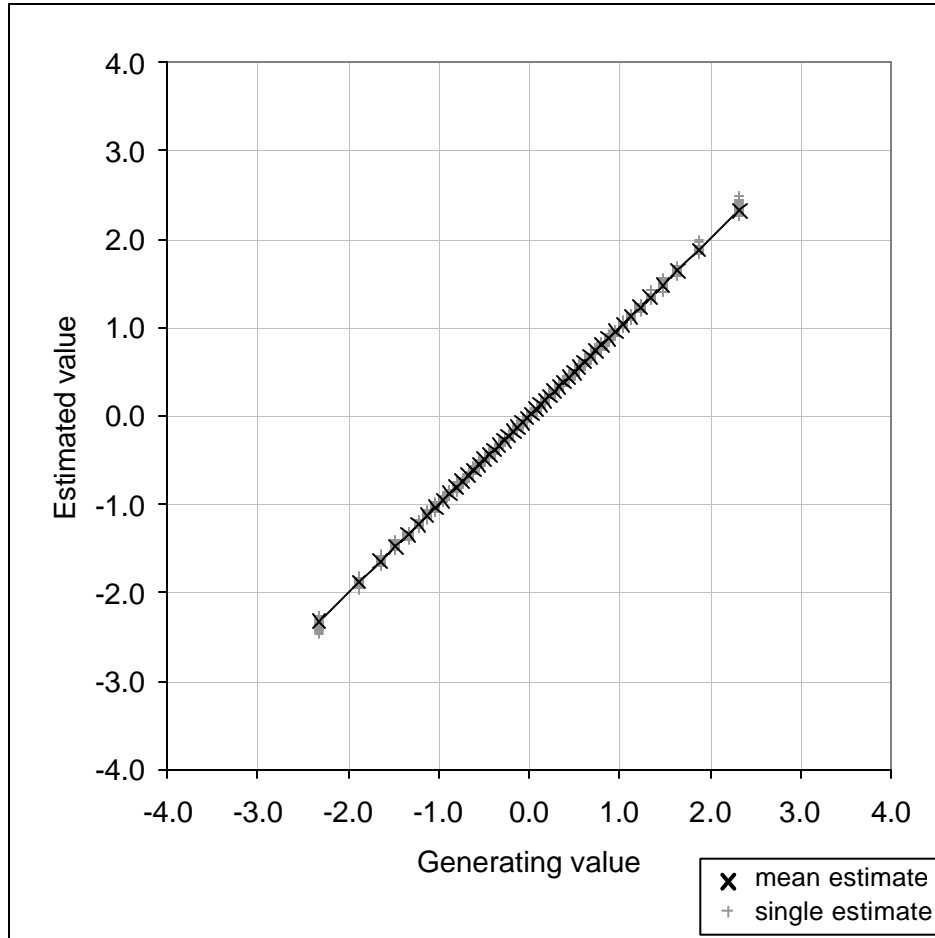


Figure 9. Comparing the standardized 2PL item difficulty estimates of 25 replications to the generating item difficulties with parallel ICCs and a large sample size ($a=U(0.8)$, $N=12500$).

Figures 10 and 11 present the combined scatterplots for the Rasch and 2PL models where the sample size is 12,500 and the standard deviation of the discrimination is 0.3 ($a=N(0.8,0.3)$). Here, the Rasch model results are very consistent, or reliable, but also very inaccurate (mean RMSD=0.26), while the 2PL results are both consistent and accurate (mean RMSD=0.029). The results demonstrated in Figure 10 are consistent with the results presented Figure 1.

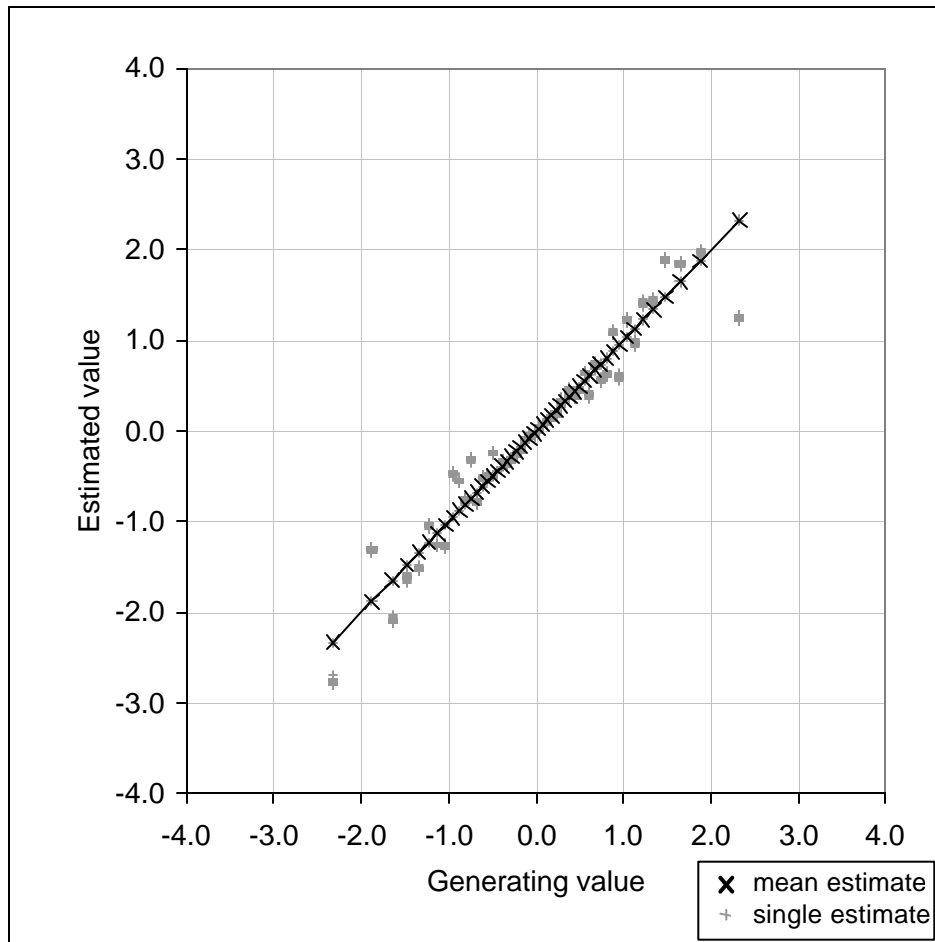


Figure 10. Comparing the standardized Rasch item difficulty estimates of 25 replications to the generating item difficulties with non-parallel ICCs and a large sample size ($a=N(0.8,0.3)$, $N=12500$).

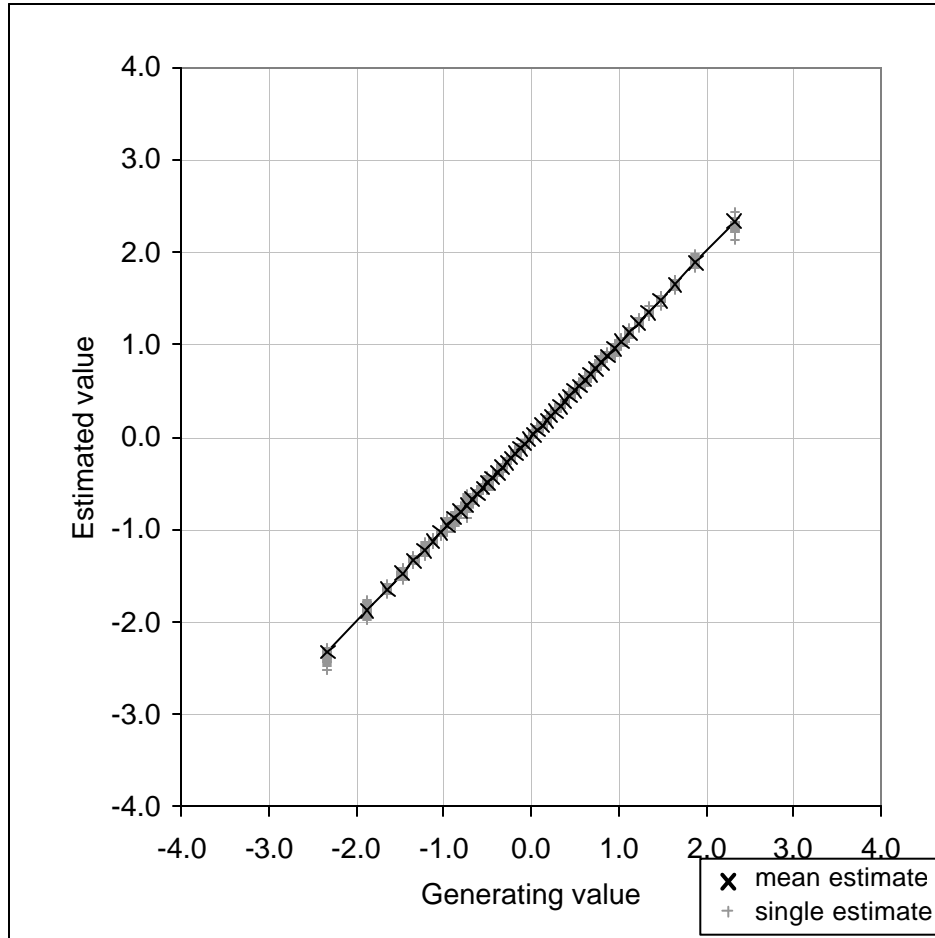


Figure 11. Comparing the standardized 2PL item difficulty estimates of 25 replications to the generating item difficulties with non-parallel ICCs and a large sample size ($a=N(0.8,0.3)$, $N=12500$).

Tables 1 and 2 present the mean RMSDs for the Rasch and 2PL estimates for the 16 sample variations (means calculated over the 25 replications). Note that the mean RMSDs from the Rasch model vary according to both the sample size and the degree of independent multidimensionality, while the 2PL model results vary primarily by the sample size. The 2PL fluctuations are likely due to the generation of different distributions of discriminations for each of the levels of independent multidimensionality (i.e. item discriminations were constant across sample sizes but were changed across each of the four levels of independent multidimensionality).

Table 1. Mean-RMSD of Rasch item difficulty estimates over 25 replications

degree of independent multidimensionality	Sample Size			
	100	500	2500	12500
0	0.19	0.08	0.04	0.02
0.1	0.19	0.10	0.08	0.07
0.2	0.27	0.20	0.19	0.19
0.3	0.30	0.27	0.26	0.26

Table 2. Mean-RMSD of 2PL item difficulty estimates over 25 replications

degree of independent multidimensionality	Sample Size			
	100	500	2500	12500
0	0.23	0.11	0.05	0.03
0.1	0.22	0.11	0.05	0.02
0.2	0.24	0.12	0.06	0.03
0.3	0.25	0.13	0.06	0.03

Figure 12 presents the estimated (using interpolated data points) Rasch and 2PL error contours along with estimated points of error contour intersection and a trendline demarking regions of model superiority. The Rasch model is superior below and to the left of the trendline. A power trendline and equation was generated automatically with the Excel trendline feature and yielded $y=1.8x^{-0.5}$. Again note that the 2PL model error contours appear to be independent of the degree of independent multidimensionality, although they are affected somewhat by the variations in the distribution of item discriminations.

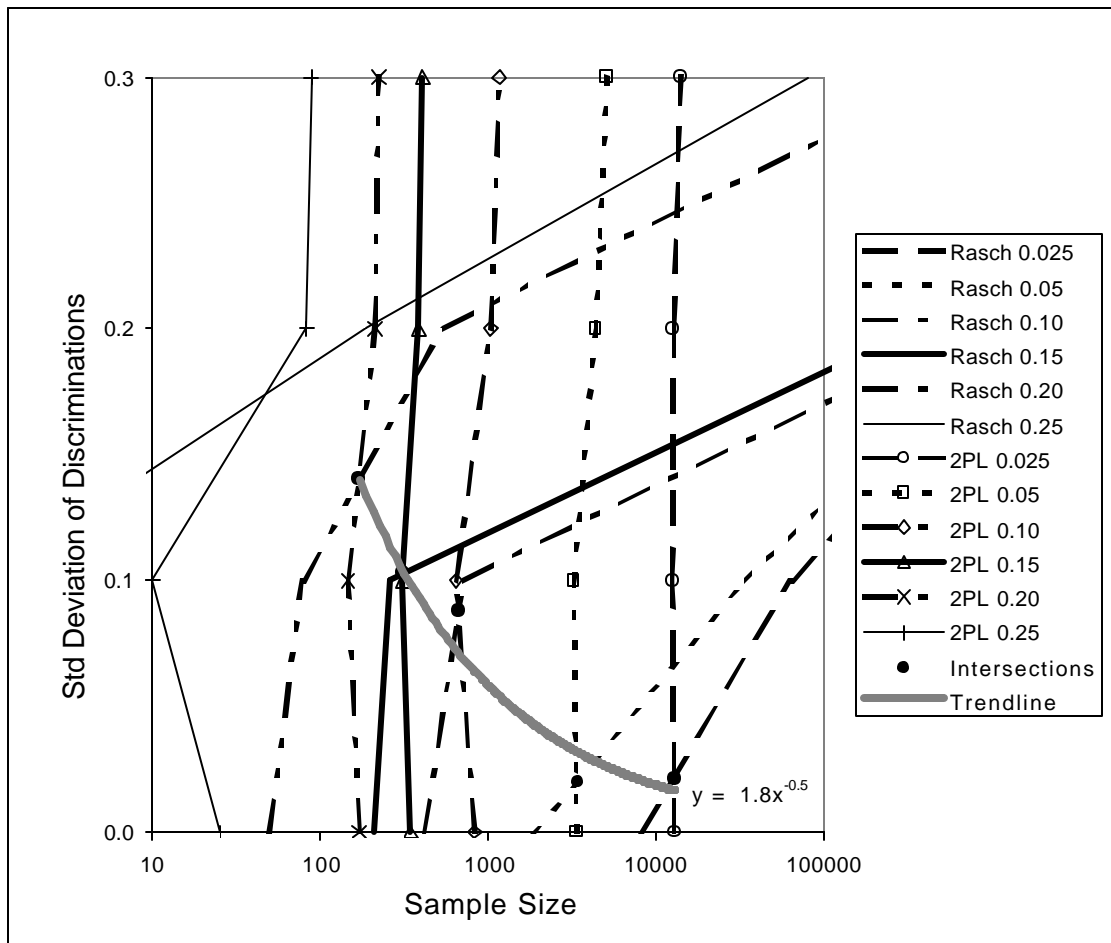


Figure 12 Combined Rasch and 2PL error contours using all items in the item set. Note the estimated points of error contour intersection and the trendline demarking regions of model superiority (the Rasch model is superior below and to the left of the trendline).

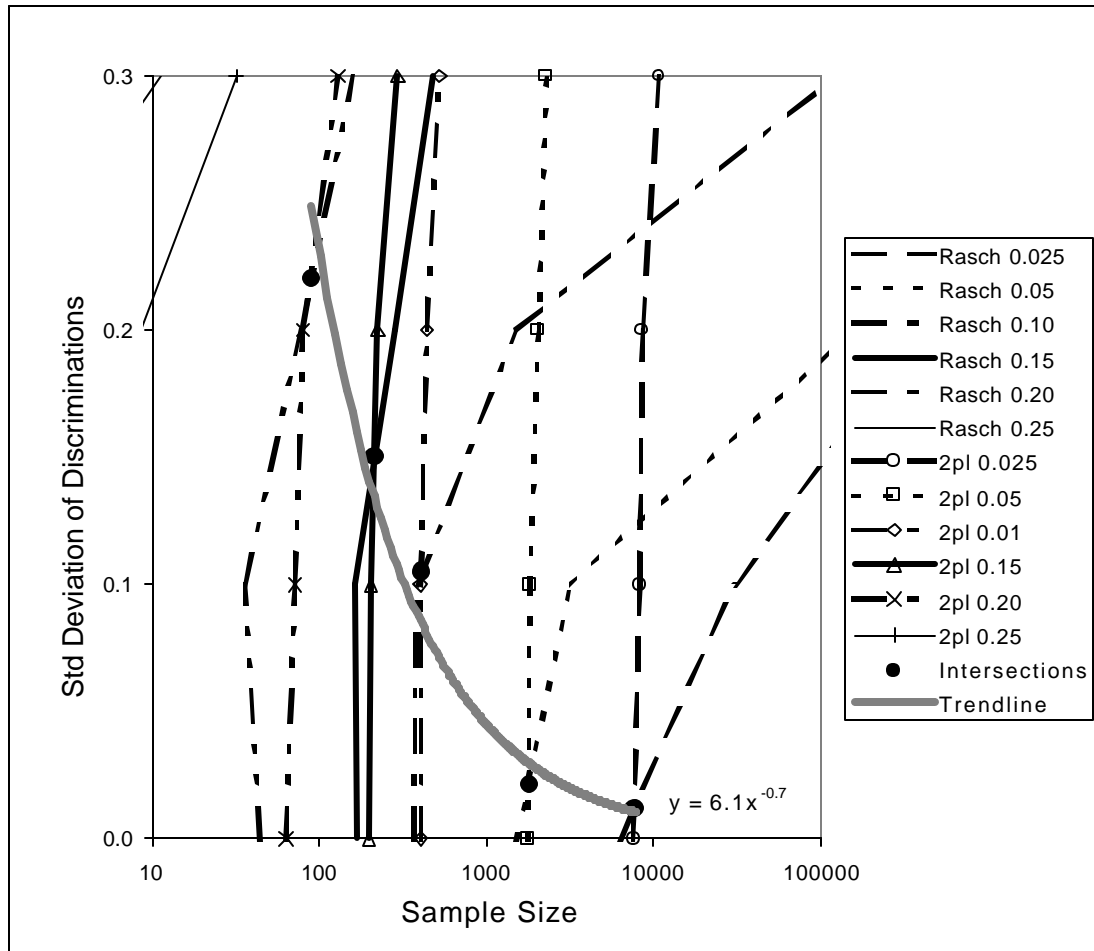


Figure 13 Combined Rasch and 2PL error contours generated using the central 60% of items in the item set. Note the estimated points of error contour intersection and the trendline that demarking regions of model superiority (Rasch model is superior below and left of trendline).

Figure 13 presents the same error contours, estimated points of error contour intersection and trendline demarking regions of model superiority as Figure 12, but using the mean RMSD from only the central 60% of the item sets to illuminate the changes in relative position of error contours and trendline when the person sample abilities more broadly cover the sample of item difficulties. In this figure it can be seen that both the Rasch model and the 2PL error contours are less affected by the distribution of item discrimination variation – resulting in more coherent error contours. Specifically, the contours have been shifted lower with respect to the sample size and higher with respect to the degree of independent multidimensionality. The automatically generated trendline was defined as $y=6x^{-0.7}$ – which on overlaying with Figure 12 shows a slight increase in the Rasch advantage with smaller samples and a

decrease in the region of Rasch superiority with larger samples. With other range restrictions (40% and 80%) the error contours improve correspondingly and the trendlines demarking regions of superiority are similar to the trendlines observed in Figures 12 and 13. When the range of items is narrowed to the central 20% of items, the error contours for the two models are essentially equivalent.

Table 3 presents the results of a simple comparison of the mean-RMSDs for the Rasch and 2PL estimates (as seen in Tables 1 and 2) for the full set of items across each of the 16 sample size by independent multidimensionality combinations (the same pattern was observed when using the central 60% of items). The relative superiority of the 2PL model over the Rasch model (or vice versa) is minimal for all small sample sizes, but becomes substantial as the sample size and the degree of independent multidimensionality increases.

Table 3. Identify superior model using Mean-RMSD for full set of items

degree of independent multidimensionality	Sample Size			
	100	500	2500	12500
0	Rasch	Rasch	Rasch	Rasch
0.1	Rasch	Rasch	2PL	2PL
0.2	2PL	2PL	2PL	2PL
0.3	2PL	2PL	2PL	2PL

Tables 4 and 5 present the mean standard deviation of the standardized InFit and OutFit statistics generated by BIGSTEPS (Rasch) over 25 replications for each of the 16 trials.

Table 4. Mean standard deviation of zstd-InFit statistics generated by BIGSTEPS (Rasch)

degree of independent multidimensionality	sample size			
	100	500	2500	12500
0	0.88	0.85	0.86	0.87
0.1	0.95	1.21	2.25	4.66
0.2	1.16	1.95	4.06	6.81
0.3	1.48	2.79	5.57	8.05

Table 5. Mean standard deviation of zstd-OutFit statistics generated by BIGSTEPS (Rasch)

degree of independent multidimensionality	sample size			
	100	500	2500	12500
0	0.93	0.91	0.95	0.94
0.1	0.95	1.22	2.13	4.32
0.2	1.13	1.84	3.81	6.64
0.3	1.44	2.64	5.26	8.12

Conclusion

As anticipated, the Rasch model yielded superior results when the underlying structure of the item set was consistent with the assumptions of the model (i.e. parallel ICCs) and when the person sample size was small (see Table 1). In addition, the Rasch model matches the accuracy of the 2PL model in cases where the item difficulty distribution is very narrow and centered well within the person sample distribution.

The evidence supports the initial conjecture (back plane of Figure 3) that the Rasch model will produce more accurate item difficulty estimates when the sample size is small and there is only a moderate amount of misfit (i.e. the Rasch model requires less information to produce useful results) and when sample size is large and there is no misfit.

When ICCs cross however, it appears that the 2PL model should often be applied. By reviewing tables 3, 4, and 5 the following, somewhat obvious, rule of thumb is suggested for use in cases where guessing is minimal:

When the standard deviations of the standardized (zstd) InFit and OutFit statistics are one or less, the use of the Rasch model is indicated, otherwise the 2PL model should be considered for the calibration of item difficulties.

This reflects the earlier observation that the theoretical advantage of the Rasch model – with respect to accurately locating items on a measurement scale – diminishes quickly as model assumptions are violated. Although it has been asserted that the existence of crossing ICCs is detrimental to meaning, it appears that meaning can be maintained when the measurement model used is able to accommodate the underlying independent multidimensionality.

Limitations and future research

The simulated data is very simplistic, using normally distributed on-target item and person samples and the 2PL model to generate simulated responses does not accurately reflect the typical complexities of true data. The use of the 2PL model for data generation, will likely have advantaged the 2PL model over the Rasch model and the procedures to correct for this advantage are not obvious. It is also unknown how dependent multidimensionality or guessing may affect item difficulty estimates in each model. Because of these concerns, the results are not general and may only provide a rough a guide.

While off-target items or persons should be expected to provide less information and appear to be detrimental to the relative accuracy of the Rasch model over the 2PL model item difficulty estimates, the examination of the effects of other person and item sample distributions might allow for further illumination of the nature of this relationship.

Misfit is currently used to modify the estimate of error in Rasch model software (e.g. Real SE, (Linacre and Wright 1999)). Given the consistency between the stability of item position estimates across replications when large samples of persons are available and the narrowing of the SE estimates (see Figure 10), it might be possible to improve Rasch item position estimates using some function of the misfit estimate and the SE estimate.

Further research is planned to complete the exploration of the conjecture presented in Figure 3 for each of the three common models: Rasch, 2PL and 3PL.

References

- Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. Statistical theories of mental test scores. F. M. Lord and M. R. Novick. Reading, MA, Addison-Wesley.
- Bond, T. G. and C. M. Fox (2001). Applying the Rasch model: Fundamental measurement in the human sciences. Mahwah NJ, Lawrence Erlbaum Associates.
- Brogden, H. E. (1977). "The Rasch model, the law of comparative judgement and additive conjoint measurement." Psychometrika **42**: 631-34.
- Cohen, M. R. and E. Nagel (1934). An introduction to logic and the scientific method. New York, Harcourt, Brace and World.
- Guttman, L. (1944). "A basis for scaling qualitative data." American Sociological Review **9**: 139-150.
- Hambleton, R. K. and R. J. Rovinelli (1986). "Assessing the dimensionality of a set of test items." Applied Psychological Measurement **10**(3): 287-302.
- Hambleton, R. K., H. Swaminathan, et al. (1991). Fundamentals of Item Response Theory. Newbury Park CA, Sage Publication.

- Hattie, J. A. (1984). "An empirical study of various indices for determining unidimensionality." Multivariate behavioral research **19**: 49-78.
- Karabatsos, G. (2001). "Understanding Rasch measurement: The Rasch model, additive conjoint measurement, and new models of probabilistic measurement theory." Journal of applied measurement, **2**(4).
- Linacre, J. M. (2000). "Item discrimination and infit mean-squares." Rasch measurement transactions **14**(2): 743.
- Linacre, J. M. (2001). "Standardized Mean-Squares." Rasch Measurement Transactions **15**(1): 813.
- Linacre, J. M. and B. Wright (1999). A user's guide to WINSTEPS, BIGSTEPS, MINISTEP-Rasch-model computer programs. Chicago, MESA Press.
- Lord, F. M. (1980). Applications of item response theory to practical testing problems. Hillsdale NJ, Lawrence Erlbaum Associates.
- Luce, R. D. and J. W. Tukey (1964). "Simultaneous conjoint measurement." Journal of Mathematical Psychology **1**: 1-27.
- McDonald, R. P. (1981). "The dimensionality of tests and items." British Journal of Mathematical and Statistical Psychology **34**: 100-117.
- McDonald, R. P. (1999). Test theory: a unified treatment. Mahwah, NJ, Lawrence Erlbaum Associates.
- Mislevy, R. J. and R. D. Bock (1990). BILOG. Mooresville, IN, Scientific Software, Inc.
- Pelton, T. (2002). Where are the limits to the Rasch advantage? International Objective Measurement Workshop, New Orleans.
- Pelton, T. and C. V. Bunderson (2002). Which measurement model is more accurate? A Monte Carlo study of the relative estimation error using unidimensional measurement models given varying amounts of multidimensionality and pseudo-guessing in the data. Large Scale Assessment in Canada, Victoria, Public Education Policy Research Group, UVic.
- Perline, R., B. D. Wright, et al. (1979). "The Rasch model as additive conjoint measurement." Applied Psychological Measurement **3**(2): 237-255.
- Rasch, G. (1960, 1980, 1993). Probabilistic models for some intelligence and attainment tests. Chicago, MESA Press.

Wright, B. (1991). "Rasch vs Birnbaum." Rasch Measurement Transactions **5**(4): 178-9.

Wright, B. D. (1999). Fundamental measurement for psychology. The new rules of measurement. S. E.

Embretson and S. L. Hershberger. Mahwah NJ, Lawrence Erlbaum Associates.