André A. Rupp

and

Bruno D. Zumbo
University of British Columbia

Which Model is Best? Robustness Properties to Justify Model Choice Among Unidimensional IRT Models under Item Parameter Drift

This article extends recent research on item parameter drift by investigating the robustness properties of basic unidimensional IRT models. Specifically, the article explores whether it is possible to advocate the consistent choice of one model over another based on its robustness properties under drift. On the one hand, it is shown that the biases that are introduced due to drift are minor for most practically relevant circumstances across all models. On the other hand, it is shown that the mathematical structure of the biases is theoretically complex so that globally superior performance of one model over another is observed only under restrictive side conditions.

Although various measurement models based on item response theory (IRT) are available to psychometric modelers (for overviews see, e.g., Junker, 1999; Rupp, 2002; van der Linden & Hambleton, 1997), the practical day-to-day applications of more complex IRT models are rather limited. Instead, it is common for testing agencies, research institutes, and consultants to use one of the more basic unidimensional IRT models. For dichotomously scored items, the one-parameter logistic (1PL) or Rasch model, the two-parameter logistic (2PL), or the three-parameter logistic (3PL) models are commonly used, whereas for polytomously scored items the graded response model (Samejima, 1969) and the multiple-response model (Thissen & Steinberg, 1984) are commonly used. In fact, some of the most popular estimation software such as BILOG-MG (Zimowski, Muraki, Mislevy, & Bock, 1996) and MULTILOG (Thissen, 1991) are designed to estimate primarily these basic powerful models.

Given some of the discrepancies between theoretical availability of models and practical model choice, it is of inherent interest to understand some of the driving forces behind model choice and to investigate whether it is possible to collect some evidence that supports the eventual model choice based on mathematical criteria about its inferential properties. This article seeks to provide some pieces of evidence that underscore that the process of model choice is complex and cannot be unambiguously resolved by invoking certain robustness properties under lack of invariance (LOI) due to item parameter drift

Bruno Zumbo is a professor of measurement, evaluation, and research methodology and a member of the Institute of Applied Mathematics and the Department of Statistics.
André Rupp is now an assistant professor in the Faculty of Education at the University of Ottawa.

(IPD). The developments in this article build on recent research in IPD, which has investigated the effect of certain types of item parameter drift on examinee parameter estimates (Wells, Subkoviak, & Serlin, 2002) as well as the effect of IPD on item response probabilities (Rupp & Zumbo, 2003). In this article, then, these investigations are extended to incorporate a cross-model comparison, and it is shown how biases due to drift of the item difficulty parameter may be only minor from a practical perspective, but their implications about model choice are rather complex from a mathematical perspective.

Consequently, the issue of model choice as viewed through this LOI lens can either be considered practically unimportant, as biases in response probabilities are relatively minor across all models; or theoretically important, as the bias patterns show that no single model possesses globally superior bias property under drift of the item difficulty parameter. But before these arguments are analytically and numerically developed, it is worthwhile to say a few words about the process of model choice in general.

Choosing a Measurement Model

At the simplest level choosing a measurement model can be a matter of training or tradition. In the former case, knowledge about a certain class of measurement models such as classical test theory (CTT) models or structural equation models (SEM) may lead one to choose one modeling approach over the other. This is, of course, a reasonable course of action, because in order to use a model responsibly, one needs to be knowledgeable, at least to a reasonable degree, about its mathematical structure, its parameter estimation process, the meaning of the different parameters, the meaning of information provided in the output, and the contexts in which it can be used. Hence if one is knowledgeable about a particular class of models, one is typically at least a responsible user of those models, which may be better than being an irresponsible user of a larger class of models that one knows little about. If model choice is driven by tradition, a similar logic often underlies the process. For example, by reputation and perception the Educational Testing Service and CTB McGraw/Hill are companies that are known often to use the 3PL IRT model whereas many language testers and the National Board of Medical Examiners often use the 1PL IRT or Rasch model. Similarly, many educational and psychological researchers still use CTT models, and the University of Iowa is often thought of as one of the centers for its cousin, generalizability theory. As in almost all areas of life, traditions have histories that are hard to break, because they have been shown to be consistently beneficial to the users and be of high practical utility.

In many practical scenarios, model choice is driven by real constraints that models place on the data input and the consumer of their output. For example, simpler models have fewer parameters that need to be estimated and thus make less stringent sample size requirements for stable parameter estimation. But even if parameters in more complex models can be estimated to a desired degree of accuracy, it may be difficult to interpret them meaningfully from a substantive theoretical viewpoint, and this may be what is desired. Conceptually, one could thus argue that the question of model choice can be answered by invoking more fundamental properties of measurement, which would, for example, favor the 1PL or Rasch model over other models because it can be considered an instance of additive conjoint measurement (Perline, Wright, &

Wainer, 1979). Researchers working seriously in cognitive assessment might not even consider a mathematical model unless it allows them formally to operationalize elements of the substantive theory as parameters in the model. For these researchers, a model of choice needs to have an appropriate mathematical structure and needs to provide adequate fit to the data and the underlying theory that generated the data.

Alternatively, one could argue that the choice between models can be considered a question of empirical model fit, which implies that every new dataset should be modeled with the model that best fits the data. However, detecting model fit for real datasets can be a tricky issue in particular for smaller datasets. Indeed, as seen above in the discussion of model complexity, sample size is often one of the main determining factors for model choice, even though sample size requirements are somewhat moderated these days by advances in the theory and implementation of Bayesian estimation paradigms (see Rupp, Dey, & Zumbo, in press, for an overview). But although it is relatively easy to find fault with a given application of a model to a dataset either on empirical or theoretical grounds, even strong critics are not always able to offer superior alternatives (see, e.g., Traub, 1983, for a passionate criticism of unidimensional IRT models).

It is thus of inherent interest to determine whether a consistent preference for a given model bears positive or negative consequences in the long run. To answer this question, this article explores under what conditions of IPD the largest amounts of bias in response probabilities occurs to determine whether one can argue for globally superior properties of one model over another. IPD, an instantiation of LOI, is a commonly observed phenomenon where item parameters from a previous calibration of an item set appear to have changed (i.e., drifted) over time. IPD has a practical effect on decision-making, because if it goes undetected, a bias in response probabilities is introduced that affects the true score estimates of examinees and, by implication, their respective θ estimates. Thus if one could make general statements that showed one model to be superior in that it resulted generally in smaller amounts of bias across different drift conditions, more substantial arguments about model choice could be developed. This article explores analytical, numerical, and visual methods to answer this question and highlight the conditions that have to hold to make such general claims. The article focuses on basic unidimensional IRT models for dichotomously scored item sets (i.e., the 1PL, 2PL, and 3PL models) as these facilitate discussions of the underlying theoretical concerns. All discussions are developed at the *population level* and hence circumvent the problem of detecting the true amount of bias in the population with calibration samples. This is not necessary for the purpose of this article, as the population analogue as a clean idealization shows the minimum and maximum amounts of biases with no confounding due to sample-to-sample fluctuation.

In an earlier study (Rupp & Zumbo, 2003), the relationships that exist between parameters from different populations on both the logit and probability scale were made explicit, and the effect of IPD on item response probabilities as well as examinee true scores was demonstrated analytically, numerically, and visually. Recall that for basic unidimensional IRT models the parameter that indicates the inflexion point of an item characteristic curve

(ICC) is often called the "item difficulty" parameter and is denoted by β_j; the parameter that indicates the slope of an ICC at its inflexion point (i.e., the "steepness" of the curve) is often called the "item discrimination" parameter and is denoted by α_j; and the parameter that indicates the lower asymptote of an ICC is often called the "item pseudo-guessing" parameter and is denoted by γ_j. Consistent with previous research (Wells et al., 2002) it was found that for moderate amounts of item discrimination parameter drift (i.e., α-drift), item difficulty parameter drift (i.e., β-drift), and joint α- and β-drift, the effect on examinee true scores was relatively minimal. The model that was used in these studies was the 2PL, and this article presents their natural extension by offering a cross-model comparison of the effect of IPD on item response probabilities and examinee true scores.

Technically, there are nine scenarios to consider for this purpose, because one is dealing with two design factors that can be crossed, (a) the generating model (3 levels: 1PL, 2PL, 3PL), and (b) the fitted model (3 levels: 1PL, 2PL, 3PL). However, certain types of parameter drift cannot be considered for all scenarios. In particular, one cannot consider α-drift for the 1PL, because by default all discrimination parameter values are held constant for this model (typically ($\alpha_j = 1$ for all items) and a "drift" for some items would effectively constitute going from a 1PL to a 2PL. In other words, it would constitute an instance of *model misfit* and not IPD, which presumes that the same model holds across calibrations for the item set under consideration. In addition, one cannot consider γ-drift for either the 1PL or the 2PL, because this would constitute another instance of model misfit as only the 3PL allows differing lower asymptote parameters across items. It is clear that item parameter drift can only be considered across models for an item parameter that is common across all models; hence one has to limit attention to the item difficulty parameter β.

The guiding idea and questions for this article are the following. Consider an item that displays β-drift and further consider that the item is calibrated with a 1Pl, 2PL, and 3PL. For what combinations of α-values, examinee locations on the latent scale, and amounts of β-drift is the bias in item response probabilities and examinee true score smallest (i.e., where are *local minima* for bias)? In addition, is it possible to identify one of these models as the one that has the smallest overall amount of bias due to the mathematical structure of the model (i.e., is there a *global minimum* of bias across models)? To answer these questions, a formalization of bias is necessary.

Formalization of Bias in Response Probabilities

To begin, define the bias in response probabilities for examinee *i* on item *j* as the difference $\Delta_{ij} = P_j(\theta_i) - P'_j(\theta'_i)$ where the first probability is computed using the original value of the item difficulty parameter β_j whereas the second probability is computed using β'_j the drifted value of the item difficulty parameter (i.e., for a drift of magnitude τ_j, $\beta'_j = \beta_j + \tau_j$). The response probability is computed using the model under consideration so that for the 3PL,

$$P_j(\theta_i) = \gamma_j + (1 - \gamma_j)\,\frac{\exp(\alpha_j(\theta_i - \beta_j))}{1 + \exp(\alpha_j(\theta_i - \beta_j))}\;;\alpha_j > 0, -\infty < \beta_j, \theta_i < \infty, 0 \leq \gamma_j < 1,$$

for the 2PL,

$$P_j(\theta_i) = \frac{\exp(\alpha_j(\theta_i - \beta_j))}{1 + \exp(\alpha_j(\theta_i - \beta_j))} \;;\alpha_j > 0, -\infty < \beta_j, \theta_i < \infty\,,$$

and for the 1PL or Rasch model,

$$P_j(\theta_i) = \frac{\exp(\theta_i - \beta_j)}{1 + \exp(\theta_i - \beta_j)} \;;-\infty < \beta_j, \theta_i < \infty\,.$$

Because the 3PL is the most general of the three unidimensional models, one can compare the difference in biases across models most easily if one uses it as a starting point. The above definition of Δ_{ij} results in the following expression of bias for the 3PL:

$$\Delta_{ij}^{3PL} = \left[\gamma_j + (1-\gamma_j)\frac{\exp(\alpha_j(\theta_i - \beta_j))}{1 + \exp(\alpha_j(\theta_i - \beta_j))}\right] - \left[\gamma_j + (1-\gamma_j)\frac{\exp(\alpha_j(\theta_i - (\beta_j + \tau_j)))}{1 + \exp(\alpha_j(\theta_i - (\beta_j + \tau_j)))}\right]$$

$$= \left[\frac{\psi_{ij}}{1+\psi_{ij}} - \frac{\psi_{ij}}{\eta_j + \psi_{ij}}\right] - \gamma_j\left[\frac{\psi_{ij}}{1+\psi_{ij}} - \frac{\psi_{ij}}{\eta_j + \psi_{ij}}\right] = (1-\gamma_j)\left[\frac{\psi_{ij}}{1+\psi_{ij}} - \frac{\psi_{ij}}{\eta_j + \psi_{ij}}\right]$$

where $\eta_j = \exp(\alpha_j\tau_j)$ and $\psi_{ij} = \exp(\alpha_j(\theta_i - \beta_j))$.

Because the 2PL is a special case of the 3PL with $\gamma_j = c$, $c \in (0,1)$, where c is typically chosen to be 0, one can analytically see that bias for the 2PL can be considered a "special case" of the bias for the 3PL under certain conditions as

$$\Delta_{ij}^{2PL} = \frac{\psi_{ij}}{1+\psi_{ij}} - \frac{\psi_{ij}}{\eta_j + \psi_{ij}}\,,$$

and hence

$$\Delta_{ij}^{3PL} = (1 - \gamma_j)\,\Delta_{ij}^{2PL}\;.$$

This equation illustrates analytically the first result of cross-model comparisons. If one calibrates an item with a 2PL and a 3PL and the item discrimination value is the same under both model calibrations, then for an identical difference between examinee and item location on the latent θ scale, the introduced bias is smaller under the 3PL. Specifically, it is smaller by a factor proportional to the probability of not guessing.

This result makes sense if one considers the ICCs for this scenario. Both ICCs would have the same slope and the same location displacement under drift; yet because the lower-asymptote for the 3PL is higher than for the 2PL, the horizontal displacement due to β-drift results in a smaller *vertical* difference Δ_{ij} between the curves for the 3PL than the identical displacement for the 2PL. It is already clear that the difference is not going to be very large because lower-asymptote parameter values are rarely larger than .3; Figure 1 illustrates this for an item with $\beta_j = 0$, $\beta'_j = .6$, $\alpha_j = 1.5$, and $\gamma_j = .3$ under the 3PL.

Because the 1PL is a special case of the 2PL and, by implication, of the 3PL, the bias under a 1PL is also a special case of the bias under a 2PL and 3PL. However, it cannot be written as compactly in an equation as above, because the difference between a 1PL and the other two models is that the discrimination parameter is fixed, which is a parameter that is not directly on the probability scale and hence precludes a simple analytic equation. This scenario is discussed further in the next section where graphics are used to illustrate these biases, which makes patterns in biases easier to describe.

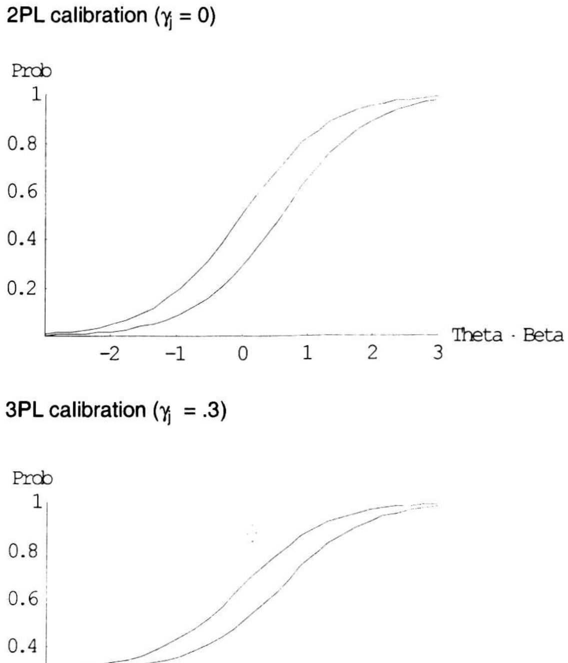
Visualization of Biases Using the Δ_{ij} Function

To appreciate the differential effect of various types of biases and their dependencies on model parameters, it is useful to visualize graphically the Δ_{ij} values. To understand the meaning of such plots, it is important to realize that Δ_{ij} is actually a *function* of four variables:

1. The discrimination parameter (α_j)
2. The location difference ($\theta_i - \beta_j$)
3. The lower-asymptote parameter (γ_j)
4. The amount of β-drift (τ_j)

One way to explore the differential effect these values have on the response probability bias is to plot the Δ_{ij} values as a function of the location difference and the discrimination parameter, which yields three-dimensional surfaces, and then to consider a matrix of these three-dimensional surfaces where the rows are defined by different values of the drift parameter τ_j and the columns are defined by different values of the lower-asymptote parameter γ_j. For illustrative purposes, attention is restricted to the following ranges of values:

1. The discrimination parameter (α_j): (0,2]
2. The location difference ($\theta_i - \beta_j$): [-3,3]


2PL calibration (γ_j = 0)

3PL calibration (γ_j = .3)

Figure 1. Effect of drift for item with identical discrimination calibrated with 2PL and 3PL.

3. The lower-asymptote parameter (γ_j): 0, .1, .2, .3
4. The amount of β-drift (τ_j)): .2, .4, .6, .8

This results in the 4 x 4 matrix of three-dimensional plots shown in Figure 2.

The plots in Figure 2 reflect the nested nature of the three unidimensional models and illustrate some general trends. First, it is apparent that the introduced bias is generally largest for the largest amount of drift (row 1 with $\tau_j = .8$) and smallest for the smallest amount of drift (row 4 with $\tau_j = .2$), which is intuitively clear. Second, it can be seen how for each amount of drift the amount of bias gets progressively less as the lower asymptote parameter increases (i.e., within a row, the surfaces are generally flatter for higher values of γ_j). Third, for a given model and fixed amount of drift (i.e., for a particular three-dimensional surface), the bias appears to be largest for items that have higher discrimination values than for items that have lower discrimination values for many location differences. Fourth, it is clear that the introduced biases are not very large in absolute magnitude across all surfaces unless there is an unusually large amount of drift. For example, if the introduced bias is .1, which is a reasonable typical value for many of the cases shown above, then 10 items with that amount of drift are required to produce a true-score difference of 1 point. This is a minute difference for most practical purposes and hence speaks well to the robustness properties of IRT models. This robustness result furthermore presumes that no other items in the data set display drift in the opposite direction, in which case effects might cancel out, leading to even less of an overall effect on examinee true scores for item sets.

For cross-model comparisons, the nested nature of the three models becomes important. First, the 2PL is a special case of the 3PL with a fixed lower-asymptote parameter (typically $\gamma_j = 0$ for all items) and hence in order to compare the biases introduced under a 2PL with those introduced under a 3PL one has to look across a row. Here one can see that the introduced bias is globally less for the 3PL for all identical location differences given that the item discrimination value is identical under both calibrations; this is shown analytically above. As an illustrative example, consider an extreme case where an item with discrimination parameter value of $\alpha_j = 1.5$ is calibrated under a 2PL (with $\gamma_j = 0$) and under a 3PL (with $\gamma_j = .3$) and displays a drift of $\tau_j = .8$. This corresponds to the two slices in the first and last three-dimensional surfaces in the first row of the above matrix as shown in Figure 3.

In accordance with Figure 1, one can see how the introduced bias is lower at all location differences for the 3PL, which would hold for other discrimination values as well. Note, however, that if the discrimination parameter values for a given item are different under the two models, or if different location differences are considered, one would have to inspect the height of the surfaces locally, which is discussed below.

Second, the 1PL is a special case of the 2PL with a fixed discrimination parameter value (typically $\alpha_j = 1$ for all items) and so comparing bias under the 1PL with that under the 2PL is tantamount to inspecting a horizontal slice from a particular three-dimensional surface in relation to the rest of the surface. For *most* location differences the introduced bias is higher if the item has a higher discrimination value under the 2PL than under the 1PL (i.e., typically an α_j greater than 1 under the 2PL); however, that it is not true for *all* location differences. As an illustrative example, consider an extreme case where an item is calibrated under a 1PL (with ($\alpha_j = 1$) and also under a 2PL (with ($\alpha_j = 2$) and displays a β-drift of $\tau_j = .8$. This corresponds to the three-dimensional surface in the upper left-hand corner of the above matrix and leads to the two horizontal slices shown in Figure 4. Here one can see that for location differences of about less than –.7 units or more than 1.5 units, more bias is introduced under the 1PL than under the 2PL, whereas for location differences between –.7 and 1.5 units the opposite is true.

Translating Theoretical Conditions into Practical Contexts

In order to make statements about the amount of bias introduced into the response probabilities and hence true scores under β-drift, several conditions had to be explicitly stated in each of the above case. Practitioners will recognize that most of these conditions appear somewhat unrealistic. For example, if an item set were calibrated once with a 2PL and once with a 3PL and both models seemed to provide adequate fit, as perhaps judged by some empirical fit statistic, then it is unlikely that a given item would have the same discrimination value under both models allowing for the global statements that were made earlier. Similarly, for a given examinee the θ estimate is likely to be different under the two models in this situation as well, which will result in different location differences relative to the item difficulty value for the two models, further complicating matters. Moreover, the amount of bias for an item calibrated under a 2PL on separate occasions is probably going to be different from the bias for the same item calibrated under a 3PL on separate occasions, adding yet another layer of complexity to the issue.

It thus appears that the answer to the question of which model leads to the most optimal properties in terms of introduced biases cannot be answered simply and is indeed complex. In an attempt to provide some closure, however, a few general considerations about the types of biases can be made. These considerations look at the biases one can expect if the population data are generated with a certain model (e.g., the 1PL) but a different model (e.g., a 2PL) is fit and deemed acceptable. Given the analytical results in this article, these synthetic descriptions should be understood as simplifications only.

Case 1: Generating Model is 1PL

If one fits a 1PL to data, one is effectively choosing one of the horizontal slices in a particular three-dimensional surface in a particular column in the above matrix. Fitting a 2PL or a 3PL to such data, which is really an example of overfitting, does not lead to differences in α or β values compared with those that are obtained under a 1PL calibration, because the assumptions of constant α values and zero γ values hold. Hence theoretically no differences in biases exist and practically any differences in biases between the different calibrations will be due to sampling fluctuations, will be extremely minor, and will have almost no practical effect on decision-making.

Case 2: Generating Model is 2PL

If one fits a 1PL to data that were originally generated with a 2PL, one forces all α values to be equal when in fact they are not. Depending on the degree of deviation from the fixed value the 1PL imposes and the true generating values as well as the location differences of examinees, one observes higher or lower

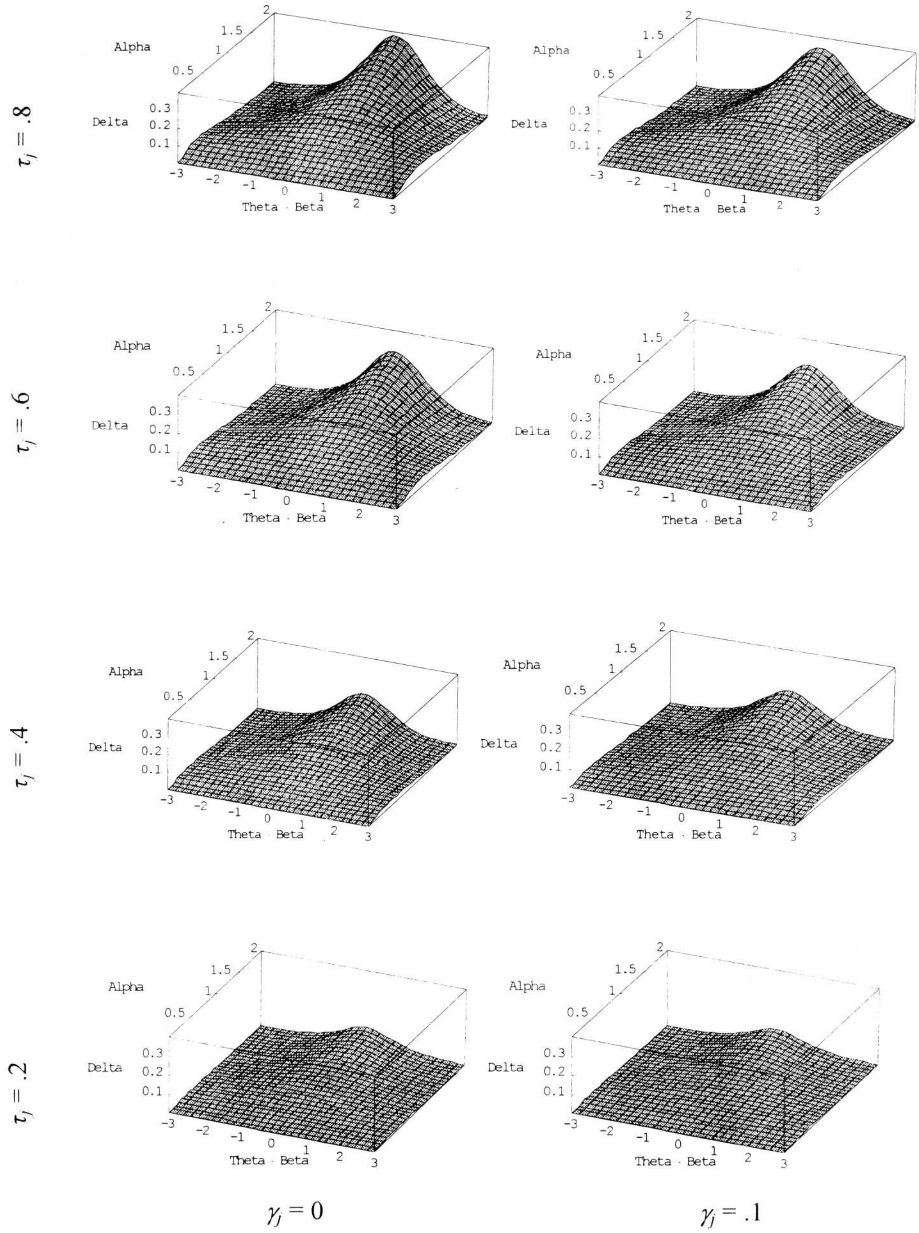
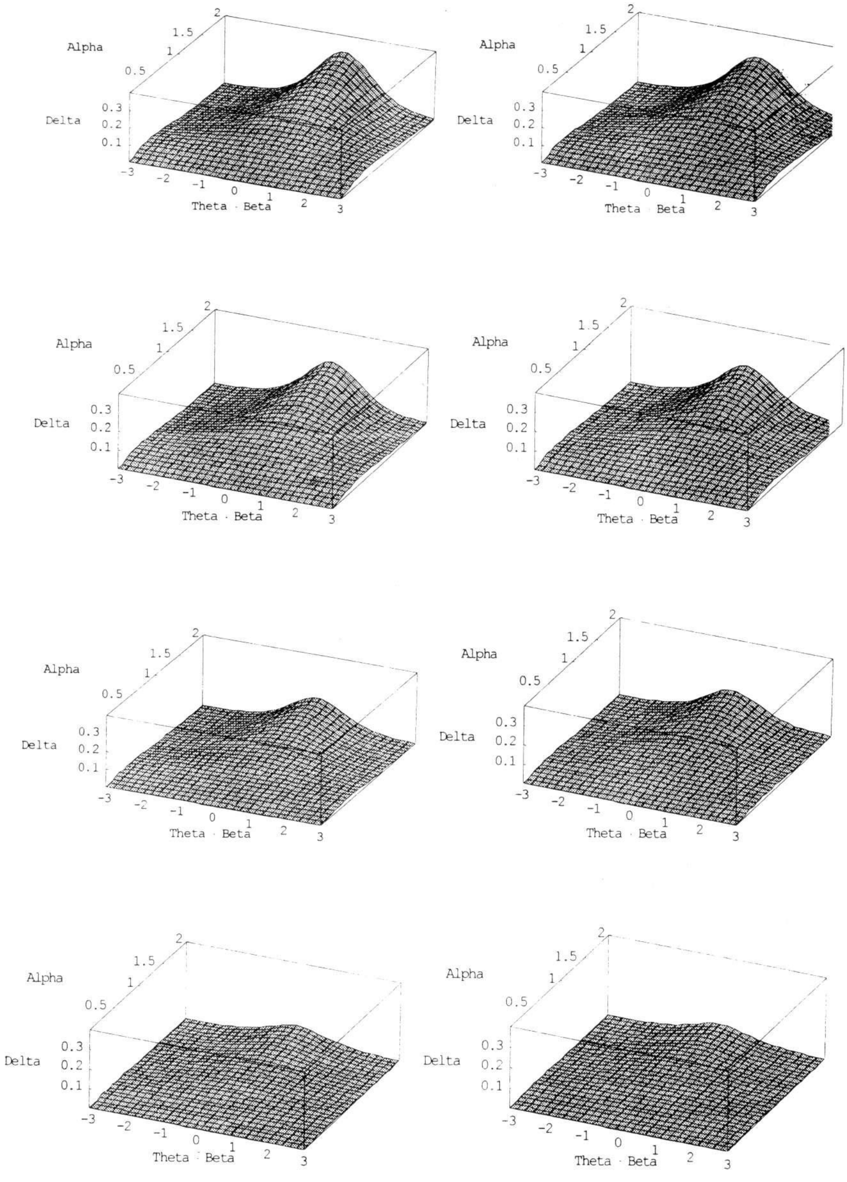


Figure 2. Surfaces for Δ_{ij} as a function of location differences, item discrimination values, and amounts of parameter drift.

biases for either model as one effectively chooses different slices of a selected surface. In extreme cases these differences might lead to some differences in decision-making, but due to potential cancellation effects across item sets these are probably going to be minor. If one fits a 3PL to data that were generated with a 2PL one will theoretically not observe any bias differences as the constant lower-asymptote assumption holds.

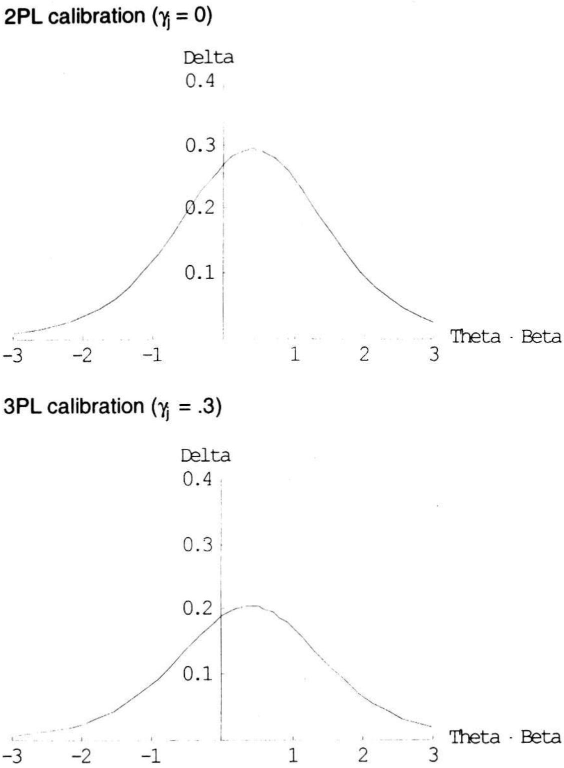


$\gamma_j = .2$

$\gamma_j = .3$

Case 3: Generating Model is 3PL

If one fits a 1PL to data that were originally generated with a 3PL, one inherits the results from the previous case, which are now compounded by the fact that one also forces the γ parameter to be 0 for all items. If one fits a 2PL to data that were generated with a 3PL, one will find that the biases under β-drift are going



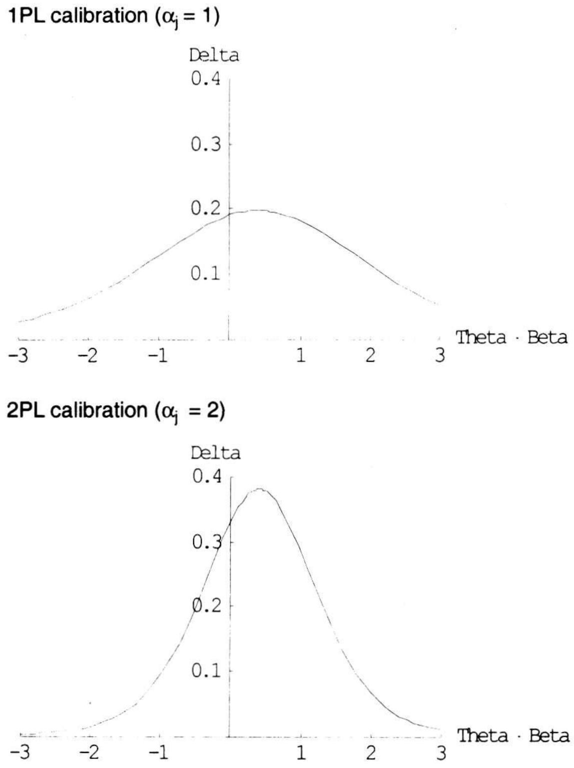
2PL calibration ($\gamma_j = 0$)

3PL calibration ($\gamma_j = .3$)

Figure 3. Effect of drift for item with identical discrimination calibrated with 2PL and 3PL.

to be uniformly larger for those items that have γ values different from 0 under the 3PL, but identical discrimination values otherwise.

Throughout this article biases in response probabilities are the focus of discussion and practitioners may wonder what the effect of such biases is on examinee parameter values. As was demonstrated in Wells et al. (2002), a *vertical* displacement in response probabilities—due to a shift in an ICC as a result of IPD-leads to a *horizontal* displacement of the associated examinee parameter values (see Figure 2, p. 80, and Figure 6, p. 85, for examples under item difficulty parameter drift). Stated differently, the overall test characteristic curve drift, which captures the cumulative effect of response probability biases from individual items on examinee true scores, can be easily translated into a related overall effect for the examinee parameters. Not surprisingly, as the authors demonstrate, minor biases in response probabilities lead to minor biases in examinee parameters (i.e., typically, the biases in examinee parameters are at most .01 units under item difficulty parameter drift). This is, of course, also true for a cross-model comparison where probability biases of similar magnitudes are observed. Assessing the exact magnitude of such effects in practice is complicated by sample-to-sample fluctuations in parameter estimates, however. This process was avoided for the purposes of this article, but the minor population effects presented in this article as an idealization of any observable real-life effects continue to speak strongly for the robustness properties of IRT models under IPD.



1PL calibration ($\alpha_j = 1$)

2PL calibration ($\alpha_j = 2$)

Figure 4. Effect of drift for item with different discrimination calibrated with 1PL and 2PL.

Conclusion

This article sets out to take an analytical look at the issue of model optimality to see how one could gather evidence for making the case that one of the three unidimensional IRT models is more robust than the others under different forms of item parameter drift, IPD. For all *practical* purposes this is not possible and it is only *theoretically* possible if one compares 2PL and 3PL calibrations under restrictive side conditions. The goal of this article is not to address calibration issues that may further compound the complexity of the problem, because it is clear that sample-to-sample fluctuations in parameter estimates affect the estimation of population biases and hence an investigator's ability to detect cleanly which case he or she is working with. At the same time, circumventing this problem is advantageous as it clarifies the underlying logic, for which it is irrelevant whether one is able to pinpoint exactly what the true parameter values are.

It is, of course, up to the modeler to decide whether robustness properties of this kind are indeed desirable. On the one hand, one may argue that IPD is a feature of the data to which a model should be sensitive so that a drift of a certain magnitude should have a strong effect on model-based inferences. On the other hand, one may argue that it is practically difficult to disentangle IPD from other factors that could lead to a differential functioning of items so that it is a nuisance that IRT models should be robust toward. Depending on the philosophical beliefs about what models are supposed to accomplish that are brought to the table by the modeler, either line of reasoning could be considered appropriate. Therefore, it is apparent that the important decision of which model to choose in which context cannot be defended on purely mathematical grounds. At the same time, it neither can nor should be defended on purely philosophical grounds, because the mathematical structure of models affects the inferences that one draws from their output. This article outlines several such effect considerations and it is hoped that the results presented here stimulate thorough and mathematically precise discussion about the factors that make model choice defensible from both theoretical and practical perspectives.

References

Junker, B.W. (1999). *Some statistical models and computational methods that may be useful for cognitively-relevant assessment.* Unpublished manuscript. Available online at http://www.stat.cmu.edu/~brian/nrc/cfa

Perline, R., Wright, B.D., & Wainer, H. (1979). The Rasch model as additive conjoint measurement. *Applied Psychological Measurement, 3,* 237-255.

Rupp, A.A. (2002). Feature selection for choosing and assembling measurement models: A building-block based organization. *International Journal of Testing, 3 & 4,* 311-360.

Rupp, A.A., & Zumbo, B.D. (2003, April). *Bias coefficients for lack of invariance in unidimensional IRT models.* Paper presented at the annual meeting of the National Council of Measurement in Education, Chicago.

Rupp, A.A., Dey, D.K, & Zumbo, B.D. (in press). To Bayes or not to Bayes, from whether to when: Applications of Bayesian methodology to item response modeling. *Structural Equation Modeling.*

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded responses. *Psychometrika Monograph Supplement, No. 17.*

Thissen, D. (1991). *MULTILOG: Multiple category item analysis and test scoring using item response theory* [Computer software]. Chicago, IL: Scientific Software International.

Thissen, D., & Steinberg, L. (1984). A response model for multiple-choice items. *Psychometrika, 49,* 501-519.

Traub, R.E. (1983). A priori considerations in choosing an item response model. In R.K. Hambleton (Ed.), *Applications of item response theory* (pp. 57-70). Vancouver, BC: Educational Research Institute of British Columbia.

van der Linden, W.J., & Hambleton, R.K. (1997). *Handbook of modern item response theory.* New York: Springer-Verlag.

Wells, C.S., Subkoviak, M.J., & Serlin, R.C. (2002). The effect of item parameter drift on examinee ability estimates. *Applied Psychological Measurement, 26,* 77-87.

Zimowski, M.F., Muraki, E., Mislevy, R.J., & Bock, R.D. (1996). *BILOG-MG: Multiple-group IRT analysis and test maintenance for binary items* [Computer software]. Chicago, IL: Scientific Software International.