A Modified IRT Model Intended to Improve Parameter Estimates under Small Sample Conditions\textsuperscript{1}

Peter Baldwin

University of Massachusetts

Abstract

This study examines the feasibility and effectiveness of grouping dichotomous test items into overlapping subsets with relatively homogeneous discriminating power and estimating subset-specific $a$-parameters for the purpose of improving item-specific parameter estimates when sample size is small. The current study only considers the 2-pl case, however, the use of Bayesian generic priors with the conventional 2-pl already provides sufficiently high-quality estimates for most small sample applications. Therefore, the purpose of this study was not to evaluate an alternate 2-pl model but rather to investigate whether the proposed strategy showed enough promise to extend to the more complicated 3-pl case. Under almost all conditions, the modified 2-pl method did as well or better than the conventional 1-pl and 2-pl IRT models. The results of this study suggest the extension to the 3-pl is warranted.
Introduction and Statement of Study Objectives

Obtaining high-quality item parameter estimates using item response theory (IRT) is a priority for many testing programs, however, this task can be challenging without large samples. At times, sufficiently large samples are simply not available. Yet, as more and more testing programs elect to administer their assessments on computers—often with on-demand testing available for examinees—item exposure concerns may result in deliberate reductions in sample sizes when large samples would otherwise be available. Regardless of the cause, the prudent psychometrician is left to do his or her best with samples that do not produce adequate parameter estimates using conventional estimation procedures.

When sample sizes are small, less general IRT models may be employed. Such models have fewer parameters, which may be adequately estimated for certain purposes with smaller sample sizes. For instance, if samples are small enough, a 1-parameter logistic (1-pl) IRT model may provide the parameter estimates with the least error compared with more general models (Lord, 1983). If it can be reasonably assumed that a given test is comprised of items with relatively homogenous discrimination parameters ($a$-parameters), the 1-pl may provide satisfactory item parameters estimates for operational use—even with small samples. However, for many assessments, this assumption is restrictive and unless items are selected for this property, $a$-parameters are not expected to be sufficiently homogenous to assume equality (Hambleton & Swaminathan, 1985). Under such conditions, new strategies are needed for modeling differences in item discrimination. The purpose of this study is to propose such a strategy and evaluate its effectiveness for improving parameter estimates under small sample conditions with particular attention given to the quality of discrimination parameter estimates.
It should be emphasized that the research presented here only represents the first step in this investigation. Specifically, the proposed strategy is evaluated only in the context of 2-pl models with small samples. For the proposed method, the 3-pl case is considerably more complex than the 2-pl case. So, while the 2-pl has fewer problems with small samples, (sensible generic Bayesian priors make the 2-pl quite usable with small samples for most applications), it was chosen here because it allows proof of concept to be investigated for the proposed method without the complications of the more general model.

**Modified IRT Models**

Methods for improving parameter estimation with small samples can be loosely organized into three categories: optimal sampling techniques, modified IRT models, and the use of auxiliary information in parameter estimation. The method proposed in this study is best categorized as a modified IRT model approach.

In modified IRT model approaches, 2- or 3-parameter models are modified in some way with the goal of improving parameter estimation under small sample conditions. In some cases these models involve novel estimation strategies, but typically they work by placing various restrictions on the possible values for one or more of the item parameters. For example, under certain conditions, some benefit may be had by fixing lower asymptotes (c-parameters) to a common non-zero value for all items (Barnes & Wise, 1991). Several researchers have proposed modified IRT model methods and evaluated their performance under small sample size conditions (e.g., Patsula & Pashley, 1996; Harwell & Janosky, 1991; Barnes & Wise, 1991; Parshall et al., 1996; Sireci, 1992; Yu & Way, 1998). Such methods are attractive to practitioners because, unlike optimal sampling strategies, they can be implemented post hoc and, unlike auxiliary information strategies, they do not require the use of collateral or other types of
information, which can be difficult or expensive to obtain (e.g., expert judgments about item
characteristics). However, gains from these approaches—in fact, for all small sample strategies
in the context of IRT—have been at best modest and with unknown generalizability. The need
for more successful small sample methods remains high.

A New Modified IRT Model

As discussed above, the 1-pl may not be suitable for many tests because of the
heterogeneity of discriminating power across items. Nevertheless, select items within such tests
may be similar enough with respect to discrimination that, were they to be grouped based on this
property, restricting within-group \(a\)-parameters to be equal would be tolerable. The approach
proposed in this study capitalizes on such within-group homogeneity of item discrimination. Just
as less general IRT models may perform better with small samples, the expectation here is that in
estimating common \(a\)-parameters for multiple items, the ratio of parameters in need of estimation
to available data is reduced, and consequently the subset-specific common \(a\)-parameters can be
reasonably estimated even with small samples. If such a strategy is to improve parameter
estimates, three important questions must be investigated; they are: (a) what measure of
discrimination is used for the purpose of forming item subsets? (b) in what manner are the item
subsets formed? and (c) how are the subset-specific discrimination parameter estimates best used
to improve item parameter estimates?

Before addressing these questions, it should be emphasized that this type of method
should be approached with a degree of healthy skepticism. Generally of concern with strategies
that attempt to extract additional information exclusively from the response data, is that they will
do nothing more than amplify the sampling errors present in the data. It remains to be seen
whether the proposed method avoids this outcome, but in its favor, collapsing parameters—in
this case by estimating a common $a$-parameter for subgroups of items—is one strategy for overcoming sampling error that is sometimes effective (e.g., collapsing score categories with polytomous IRT models).

**Overview of Procedure**

Before discussing the specifics of the study methodology, it may be helpful to present an overview of the implementation of the proposed modified IRT model. Given some test comprised of dichotomous items with item responses from $N$ examinees (where $N$ is some small number), the discriminating power of each item is estimated. This can be done using various measures; the proposed study evaluates three: biserial correlations, corrected point biserials, and 2-pl $a$-parameter estimates. Items are then rank ordered by discriminating power—i.e., item 1 has the lowest discrimination estimate, item 2 has the second lowest estimate, and so on. Next, adjacent items (based on the rank ordering) are grouped into overlapping subsets. This can be accomplished using various strategies, the simplest perhaps being to form equal-sized subsets. For purely illustrative purposes, Figure 1 shows this scenario for a 10-item test with 5-item subsets.

**Figure 1.** After rank ordering items based on their estimated discriminating power, items can then be grouped into overlapping subsets of items with relatively homogeneous discrimination.

If biserials were used as the measure of discrimination in this example, the first subset would be comprised of items 1 to 5—i.e., the 5 items with the 5 lowest biserials; the next subset would be comprised of items 2 to 6 and so on.
Grouping items in this fashion results in a new—and much longer—test, albeit one generated strictly for the purpose of item calibration. This extended test is then calibrated, restricting within-group \( a \)-parameters to be equal and restricting \( b \)-parameters associated with multiple instances of the same item to be equal (alternately, \( b \)-parameters can be fixed to a prior estimate—e.g., their corresponding 1-pl estimate). In the example above, the calibration provides 6 discrimination estimates, one for each subset \( (a_{s1}, a_{s2}, \ldots, a_{s6}) \). These 6 estimates can be used to generate the 10 item discrimination estimates using various strategies. For example, for a given item, the discrimination estimates associated with each subset in which the item appears could be averaged. Or, to the extent possible, each subset discrimination estimate could be assigned to the center-most item in each subset. Both of these strategies leave the most and least discriminating items somewhat neglected, at least relative to the middle-discriminating items. One adjustment that may ameliorate this neglect involves including smaller item subsets for those relatively high and relatively low discriminating items. Figure 2 shows the above example (from Figure 1) modified in this fashion.

**Figure 2.** Including additional, reduced-size item subsets is one way to potentially improve item discrimination estimates for relatively high and relatively low discriminating items.

<table>
<thead>
<tr>
<th>1 2 3</th>
<th>1 2 3 4 5</th>
<th>2 3 4 5 6 7</th>
<th>3 4 5 6 7 8 9</th>
<th>5 6 7 8 9 10</th>
<th>6 7 8 9 10</th>
</tr>
</thead>
<tbody>
<tr>
<td>( a_{s1} )</td>
<td>( a_{s2} )</td>
<td>( a_{s3} )</td>
<td>( a_{s4} )</td>
<td>( a_{s5} )</td>
<td>( a_{s6} )</td>
</tr>
</tbody>
</table>

Various grouping strategies and various strategies for estimating item discrimination based on subset-specific discrimination estimates are examined as part of this study. Under some conditions, results are improved by recalibrating the test using the modified model estimates as item-specific prior means.
It may be helpful to summarize this procedure with an outline of the steps needed for its implementation.

1. Estimate the discriminating power of each item. Discrimination can be estimated in many different ways. This study evaluates three: biserials, corrected point biserials and (conventional) 2-pl a-parameter estimates.

2. Rank order items based on their discrimination estimates. Rank ordering produces groups of adjacent items with relatively homogeneous discriminating power.

3. Group items. This too can be accomplished in numerous ways. In this study, the two approaches described above (i.e., equal-sized subsets and equal-sized subsets with reduced size subsets for high or low discriminating items) are evaluated. Additionally, group size is varied as a study condition.

4. Calibrate items using the 2-pl model modified such that within-subset a-parameters are constrained to be equal. Additionally, b-parameters associated with multiple instances of the same item are constrained to be equal. Item parameters can be constrained in this way using MULTILOG (Thissen, 2003).

5. Derive item discrimination estimates from the subset discrimination estimates. Two strategies are evaluated as part of this study: averaging the a-parameter estimate for each subset in which a given item appears; and to the extent possible, assigning the subset a-parameter estimate to the center most item in each subset.

**Overview of Study Method**

The procedure was evaluated using both simulated and real response data. In the case of the real data, a large data set \(N = 25,000\) from a large-scale high-stakes mathematics assessment with minimal guessing was used. Taking the full data set, the test was calibrated with MULTILOG (Thissen, 2003) using the 2-pl IRT model, and the resulting large sample estimates served as the criteria for evaluating the small sample parameter estimates. Small samples were
drawn without replacement from the total examinee pool. In the case of simulated data, the true parameter values were the criteria for evaluating the estimates. The simulated item and person parameters were based on the estimates obtained using the entire sample of real data.

The study conditions, which are summarized in Table 1, included: 3 test lengths (25, 40, and 50: a short test, a medium-length test, and a long test, respectively), 9 examinee sample sizes, ranging from \( N=100 \) to \( N=500 \) in increments of 50, 5 rank ordering criteria (random order, criterion order—i.e., large sample or true discrimination, biserial, corrected point biserials, and conventional 2-pl \( a \)-parameter estimates), 2 grouping strategies, and 50 replications. Additionally, the number of items per subset (into which similarly discriminating items will be grouped) included 3, 5, 7, 9, and 11.

<table>
<thead>
<tr>
<th>Table 1. Summary of Study Conditions</th>
</tr>
</thead>
<tbody>
<tr>
<td>Study Condition</td>
</tr>
<tr>
<td>Data Sets</td>
</tr>
<tr>
<td>Test Length</td>
</tr>
<tr>
<td>Examinee Sample Size</td>
</tr>
<tr>
<td>Item Rank Ordering Criterion</td>
</tr>
<tr>
<td>Grouping Strategy</td>
</tr>
<tr>
<td>Size of Item Subsets</td>
</tr>
<tr>
<td>Replications</td>
</tr>
</tbody>
</table>

\[ \text{Total} \quad 135,000 \]
Evaluation Criteria

Parameter estimates from the modified 2-pl model, the conventional 1-pl model, and the conventional 2-pl model were compared against 1) large sample estimates, in the case of the real data set, and 2) true parameters, in the case of the simulated data set. All calibrations were done both with generic Bayesian priors of $b \sim N(0,2)$ and $a \sim N(1,.5)$ and without any priors. All parameter estimates for both the criterion and the small sample runs were transformed using the mean-sigma transformation procedure such that for each calibration difficulty parameters had a mean of 0.0 and a standard deviation of 1.0.

Both $a$-parameters and $b$-parameters were evaluated with respect to random error and systematic error. Random error was measured by computing the mean absolute error and RMSE across replications for each item individually and then averaging the results across the items for each test. Systematic error was measured by computing the mean absolute bias for each test. This value was computed by taking the average signed error (bias) across replications for each item individually. Then, the absolute value of each item-specific bias estimate was averaged across all items for a given test. All results are reported on the normal metric.

Results

Because of the large number of study conditions, only select results are presented here. With respect to data set, the same trends were observed for both real and simulated data; therefore, only the findings using simulated data are discussed here. With respect to ranking criterion, ranking items by their conventional 2-pl estimates performed better than ranking by biserials or corrected point biserials and therefore results using these classical statistics are not presented. With respect to grouping strategy, only the results based on grouping items into equal-
size subsets of 5 items with additional reduced size subsets for the 3 most and 3 least discriminating items are presented below (see Figure 2 for an example of this grouping strategy). This approach performed better than the other grouping methods. Finally, for the modified 2-pl $a$-parameter estimates, the results that follow are based on assigning each subset-specific $a$-parameter estimate to the center item in the subset (to the extent possible). This yielded the best performance among the modified 2-pl results.

As mentioned, in most situations using generic priors is an effective means of reducing error in the 2-pl case. Therefore, it may be informative to first consider the results obtained without using generic priors to avoid confounding any gains of the modified method with gains due to the generic priors. These results were encouraging: the modified 2-pl reduced systematic error in the discrimination parameters for all sample sizes and all test lengths compared to the conventional 1- and 2-pl. With respect to absolute error in the $a$-parameters, the modified 2-pl performed better than the conventional 1-pl for all conditions except for the 40- and 50-items tests with $N = 100$ and it outperformed the conventional 2-pl for all conditions. The findings were similar for RMSE, where the modified 2-pl performed better than the conventional IRT models except in the case of $N \leq 150$, where the 1-pl performed best for the 40- and 50-item tests and in the case of the 25-item test with $N = 100$, where, again, the conventional 1-pl performed better. The conventional 2-pl always performed least well with respect to both systematic and random error for all test lengths and samples sizes. (It may be worth noting that if the $b$-parameters are fixed to their corresponding 1-pl estimates during the modified calibrations, the modified 2-pl method outperforms the 1-pl for all sample sizes and test lengths for both random and systematic error, although the magnitude of the gains overall are slightly less under these
conditions.) Figures 3 and 4 show the mean absolute bias and the mean absolute error in the a-parameter estimates, respectively, for the 40-item test. Included are the modified 2-pl results obtained by ranking based on the true values and ranking randomly as points of reference. These results suggest that the modified model is behaving in a manner consistent with expectation. When items are ranked based on their true discrimination, estimation error is smallest. When items are randomly ordered, the results are similar to the 1-pl results. Ranking based on the conventional 2-pl a-parameter estimates (or, for that matter, biserials or corrected point biserials, although these results are not shown), performs in between.

The modified method did well with the b-parameter estimates as well, although the conventional 2-pl provided the least bias for sample sizes \( N \geq 400 \). Figures 5 and 6 show these results for the 40-item test.

![Figure 5. Mean Absolute Bias in b-Parameter Estimates.](image)
Again, these findings are consistent with expectation. In most cases, the modified 2-pl that ranked items based on their true discrimination performed best, followed by the modified 2-pl that ranked items based on the conventional 2-pl estimates. When sample size exceeded 350, the conventional 2-pl also performed well.

The results presented so far offer some evidence that the modified 2-pl method is working as expected. However, in practice, one would almost certainly use generic priors on the item parameters when sample sizes are small. Therefore, having demonstrated some basic proof of concept for the proposed modified model, the next set of results will report the findings when the same data are calibrated with generic Bayesian priors of $b \sim N(0, 2)$ and $a \sim N(1,.5)$.

Using the modified model estimates as means for item-specific priors (generic prior standard deviation of about .20 seemed to work best) performed better than simply using the
modified model estimates themselves as was done for Figures 3-6. Therefore, the next set of results employ this strategy.

When generic priors are used, the conventional 2-pl and the modified 2-pl that used the 2-pl $a$-parameter estimates to rank items performed almost identically with respect to random and systematic error for both $a$- and $b$-parameter estimates. The modified 2-pl that used the true $a$-parameters to rank items performed best, although gains were extremely small with respect to mean absolute bias in the $a$- and $b$-parameter estimates. Figures 7 through 10 show these findings for the 40-item test.

---

Footnote:

2 Although using the modified 2-pl estimates as means of item-specific prior distributions improved estimates, these results were not shown in Figures 3-6. The reason for this was that using any reasonable prior (item-specific or not) has some benefit and for Figures 3-6 it was desirable to avoid confounding any such gains with those due to modified 2-pl method itself.
Discussion

The proposed modified method appears to have some merit: it performed well relative to the conventional 1-pl and 2-pl models without generic priors. Considerable differences were observed between the modified 2-pl that ranked items based on the true item discrimination and the modified 2-pl that randomly ordered items. Ranking based on the conventional 2-pl estimates performed in between these extremes, although results were typically much closer to the modified 2-pl that ranked by true $a$-parameters.

When generic priors were used, which is more in line with common practice, the modified and conventional 2-pl performed about the same, although some gains, notably with respect to random error, were evident when items were ranked by their true $a$-parameters. Of course, true values would not be known in practice, but it does suggest that even with generic priors, the modified 2-pl is still functioning as expected, even if the rankings available in practice
are not of sufficient quality to add gains above and beyond those due to the presence of generic priors.

As expected, the modified 2-pl results are not compelling enough to justify the inconvenience of the additional steps needed to calibrate the data using the proposed strategy. When sensible generic priors are employed, the conventional 2-pl does not require improvement for many purposes. Indeed, with mean absolute bias in item parameters typically less than .03, and absolute error typically less than .15, there was very little room for improvement with these data. As noted above, the small sample problem is largely a 3-pl problem. Because of this, the goal of this study was limited to the evaluation of the reasonableness of the modified method in order to determined whether implementing it with a 3-pl model was worth investigating. These results suggest that it is.

Limitations and Areas for Further Research

Early in this paper, three logistical questions were posed relating to the implementation of the proposed method:

1. What measure of discrimination is used for the purpose of forming item subsets?
2. In what manner are the item subsets formed?
3. How are the subset-specific discrimination parameter estimates best used to improve item parameter estimates?

This study provided several possible answers to these questions, but many more exist that need investigation.

For example, items were rank ordered based on three measures of discrimination: biserials, corrected point biserials, and conventional 2-pl $a$-parameter estimates. These measures have the benefit of being familiar and readily available, but may not be the best measures of
discrimination for this application. Perhaps there are measures that perform better when the only goal is rank ordering items by discrimination.

Of course, with respect to forming item subsets, rank ordering may not even be the best approach. The method may benefit more from utilizing the interval nature of these discrimination measures rather than attempting to improve the ordinal values. Interval measures of discrimination allow for more sophisticated clustering strategies than were used in the current study and such strategies could produce groups with greater homogeneity of discrimination.

With respect to translating the subset-specific discrimination parameter estimates into item-specific discrimination estimates, assigning the subset $a$-parameter to the center item in each subset performed better than averaging every subset-specific estimate for every subset in which an item appears. However, using the center item assumes that the distribution of $a$-parameters for each subset is symmetric about the center item—an assumption that certainly would not hold true for most subsets. For these data, averaging performed worse, however, perhaps a weighted average would perform better—after all, there is no reason to expect the dispersion of discrimination within subsets to be consistent.

These potential methodological refinements suggest that further research could improve the effectiveness of the procedure. However, given that there is little to be gained from the proposed strategy in the context of 2-parameter models, the initial focus of future research must be on extending the modified model to include a non-zero lower asymptote.

**Concluding Remarks**

This study examined the feasibility and effectiveness of grouping items into overlapping subsets with relatively homogeneous discriminating power and estimating subset-specific $a$-parameters for the purpose of improving *item-specific* parameter estimates when sample size is
small. Although only the 2-pl case was considered here, the goal was not to offer an alternate 2-pl model but rather to investigate whether the proposed strategy showed enough promise to extend to the more complicated 3-pl case. Small sample estimation is largely a 3-pl problem—one that is not remedied by current methods. There is much to be gained if a generalizable solution to this problem is developed.

The proposed method performed well. Under almost all conditions, the modified 2-pl method did as well or better than the conventional IRT models. Extension to the 3-pl seems warranted.
References


