THE UNIVERSITY OF IOWA

ITP

Iowa Testing Programs

Validity of the three parameter item response theory model for field test data
ITP Research Series

Brandon LeBeau
Aaron McVay

ITP Research Series

2017.1

Validity of the three parameter logistic item response theory model for field test data

Abstract

Item response theory is a large sample procedure to estimate item parameters based on individual response strings. However, what happens when the data available to estimate item parameters is small? This situation is common when new assessment items are tried out for inclusion in future operational assessments, commonly called field testing. In field tests, many items are spread out over a fixed set of respondents which can limit the number of responses on a given field test item. Four models are compared with real world field test data to evaluate their ability to accurately estimate item parameters in order to inform test developers. Implications for the four models on estimating field test item parameters are discussed.

Birnbaum's three parameter logistic item response theory (3PL IRT) model is a widely used model for assessment data (Birnbaum, 1968). The wide use of this model stems from the flexibility this model offers. The 3PL IRT explicitly allows for the test questions (commonly called items) to have varying discrimination parameter estimates across the items and also account for the nonzero likelihood of answering the item correctly by guessing (De Ayala, 2013). In addition, the model flexibility can be shown by the better fit compared to simpler models such as the two parameter (2PL) or Rasch model (CTB/McGraw Hill, 2008) that fix some of the IRT parameters to specific values (i.e. zero of one) instead of estimating them.

However, a limitation of the 3PL IRT model is the larger sample sizes required to estimate the three parameters for every item (De Ayala, 2013). This limitation can increase the uncertainty in the parameter estimates, particularly for the pseudo-guessing parameter, in small sample size conditions (Thissen & Wainer, 1982). For operational forms, sample size is commonly not a concern. However, when trying out new items for inclusion in a future operational forms, commonly called field testing or tryouts, the sample size can become much smaller where uncertainty in parameter estimates may be a problem.

Field tests are used by test developers to gather empirical data on how the new items behave when answered by an individual. The empirical data obtained can help to inform how well distractors are performing (e.g. a distractor may not be attractive to respondents), how difficult the item is, or help spot other inconsistencies. This empirical data obtained from a field test can then be used to inform which items will appear on an operational form and which items need to be edited or abandoned entirely. As a result, many more items are tried out than will appear on a final operational form. This design with many items being tried out across a fixed group of respondents has implications for the psychometric analysis; more specifically, the number of responses for a given item are commonly smaller compared to data available from an operational assessment.

This can put a strain on the psychometric analysis of these response strings when using a 3PL IRT model and increase the uncertainty of item parameter estimates (Thissen & Wainer, 1982). With psychometric analyses aiding in the selection of which items appear on an operational form, the increased amount of uncertainty in the parameter estimates can devalue the information available to those making these decisions. This study aims to explore the validity of the 3PL model and three 3PL alternatives for calibration of field test data with small sample sizes.

Field Test Designs

As mentioned above, the goal of field testing is to tryout items before being placed on an operational form. Many standardized testing agencies employ item writers who are content experts to write new items that match content with specific standards to be evaluated on a test. As can be imagined, there are many more items written than actually make it onto an operational assessment form (Downing & Haladyna, 2006). This has a few implications, first, that many more items than are needed are tried out, and secondly, these items are spread out over a fixed set of individuals commonly resulting in a smaller number of responses for each field test item.

There are many ways to gather data on field test items. One way is to embed items directly in the middle of the operational assessment. This has become more popular as standardized tests have moved away from paper and pencil tests to being taken on a computer (both with adaptive and fixed form administrations). Embedding items has a few advantages, namely that the individual does not know which items are operational or field test; therefore they may try equally hard on all the items. This may also allow for a larger sample size depending on the specifications regarding how many items can be embedded in a single administration. However, there are drawbacks. Fatigue may be an issue if many field test items are embedded within the operational assessment and have an impact on student performance toward the end of the test. Related to fatigue, the time to take the exam may increase with embedded field test items (Downing & Haladyna, 2006).

A second popular way to conduct field tests is through creating a mini-test or booklet that contains a group of field test items. These would then be administered together, likely after the individual has completed the operational assessment. This has the advantage of not fatiguing individuals when taking the operational assessment, however has the disadvantage of the individuals now knowing these are field test items. This may impact the response strings due to individuals not giving 100% effort (Downing & Haladyna, 2006). In addition, screening for response strings where individuals did not attempt each question would need to be performed. For example, an individual may have marked 'c' for all items. This may be a response string to remove due to not adequately attempting each item, and would further reduce potential sample size constraints already present from the field test design. Secondly, depending on the pool of field test individuals, the sample size may be smaller for some field test booklets. This could occur due to being tried out on a smaller population, perhaps the group ran out of time to administer the field test booklets, or other field test administration problems that may arise.

Limitations of Field Test Design

The primary objective when conducting field tests is to gather item statistics to make informed decisions regarding whether the item will appear on an operational form. The statistics relevant for such a decision would include the proportion of respondents answering the item correctly, biserial correlations, and increasingly over the last few decades to include IRT item parameters (Downing & Haladyna, 2006).

The data available to produce the desired item statistics for field test items is commonly smaller compared to an operational form due to more items spread across a fixed pool of individuals. This has direct implication for the estimation of the IRT item parameters; especially when a 3PL model is used for estimating the IRT item parameters (a process called calibration). The strengths and weaknesses of this model are presented below for field test data (or more generally for small samples) and alternative models are considered as options to alleviate concerns over estimation with the 3PL model.

The Model

Birnbaum's 3PL model can be represented as:

$$p(x_j = 1|\theta, a_j, b_j, c_j) = c_j + (1 - c_j)\frac{1}{1 + e^{-Da_j(\theta - b_j)}}. \tag{1}$$

The equation above models the likelihood of a respondent correctly answering item j given the respondent's ability (θ), the item's difficulty (b_j), the item's discrimination (a_j), and the pseudo-guessing parameter of the item (c_j). The equation also includes a scaling constant D, which is commonly set to $D = 1$ or $D = 1.7$ depending on if the logistic or normal ogive metric is desired. The three item parameters make up the item response function which is a mathematical representation of the likelihood a respondent answers an item correctly given their ability. See De Ayala (2013) and Lord (1980) for a more thorough introduction to IRT models.

From an individual's response string, common statistical software such as BILOG-MG (Mislevy & Bock, 1990) or PARSCALE (Muraki & Bock, 1997) are used to estimate the three item parameters simultaneously using an iterative procedure while assuming a distributional form (commonly standard normal) for the ability estimates (Baker & Kim, 2004; Kolen & Brennan, 2014). This procedure is commonly called marginal maximum likelihood (Baker & Kim, 2004; De Ayala, 2013). Maximum likelihood is an asymptotic estimation procedure, therefore large sample sizes are preferred over small sample sizes to ensure the global maximum in the likelihood function is achieved. Although no definitive rules are found in the literature, general rule of thumbs state that sample sizes of at least 1000 are needed for the 3PL model, but larger samples would be preferred (De Ayala, 2013). For a more thorough discussion of IRT estimation techniques, see Baker and Kim (2004) and De Ayala (2013).

Implications of small samples on parameter estimates

The IRT simulation literature has shown that the 3PL IRT model requires larger samples for adequate estimation of the item parameters and individual ability (De Ayala, 2013; Hulin, Lissak, & Drasgow, 1982; Lord, 1980). It is often cited that approximately 1000 individuals are needed for adequate estimation (De Ayala, 2013; Hulin et al., 1982). Hulin et al. (1982) found that there is a trade-off between sample size and test length where increasing the test length improves estimation of the item parameters. Not unsurprisingly, the 2PL IRT model does not require as large of sample sizes compared to the 3PL IRT model. Parameter recovery occurs with samples as small as 500 for the 2PL model (Hulin et al., 1982).

A by-product of this is the complexity of estimating three item parameters simultaneously from the response data made up of a collection of zeros and ones. The difficulty in estimation can be directly assessed through the evaluation of the standard errors of the item parameter estimates (Lord, 1980). Prior research by Thissen and Wainer (1982) has shown that the pseudo-guessing parameter standard errors are particularly sensitive.

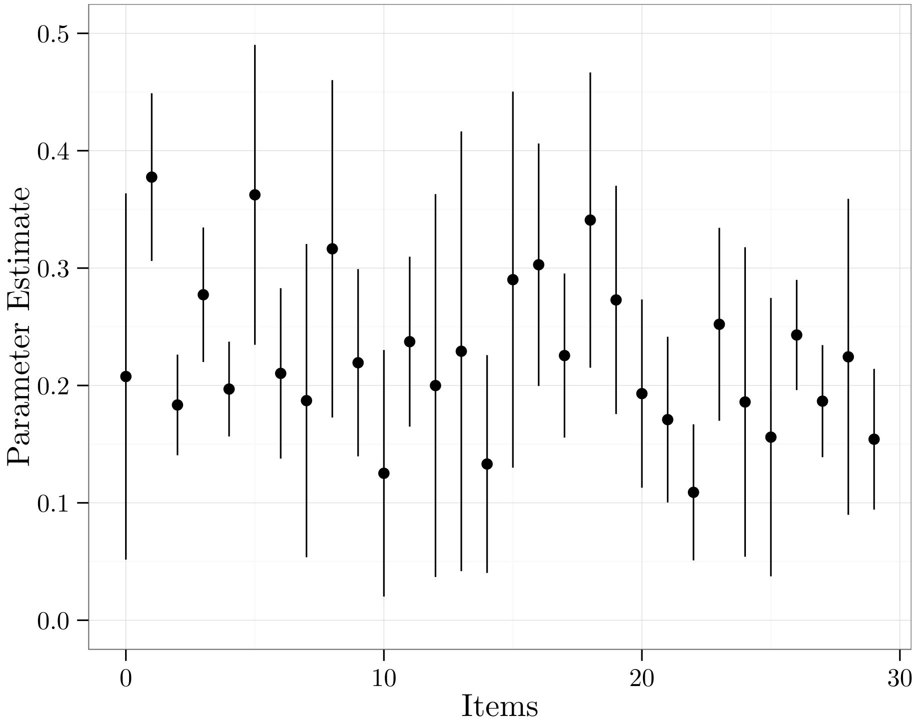


Figure 1: Pseudo-guessing estimates with 95% confidence intervals

Figure 1 shows 95% confidence intervals of the pseudo-guessing parameter for 29 field test items. The figure shows the standard errors for the pseudo-guessing parameters are quite large, on average about 0.05. This is particularly troubling when envisioning the range of the pseudo-guessing parameters from 0 to 1, but more commonly restricted by the software to be less than 0.5. If the average standard error is 0.05, then a 95% confidence interval would encompass about 40% of the common range for the pseudo-guessing parameters. This seems too large, especially when considering how changes in the pseudo-guessing parameter estimate can impact other parameter estimates. Also, the pseudo-guessing parameter estimate can drastically change the shape of the item characteristic curve for an individual item. If the estimation is poor and consistent in a single direction (i.e. overestimated) across many items this may also impact the test characteristic curve which could directly impact student scores.

Furthermore, the uncertainty found in estimating the pseudo-guessing parameters can have a direct impact on the estimates and standard errors of the discrimination and difficulty parameters (Thissen & Wainer, 1982). Figure 2 shows three item characteristic curves (ICC) for a single field test item, where the middle ICC is based on the 3PL parameter estimates for this item. The other two ICCs are adjusted based on the upper and lower limits for a 95% confidence interval for the pseudo-guessing parameter.

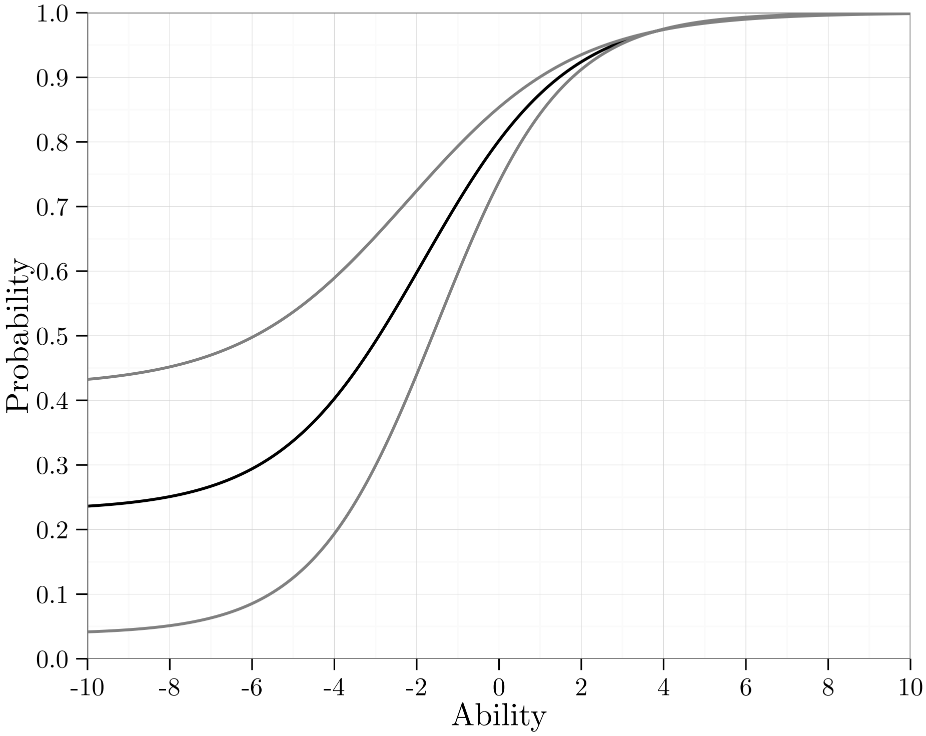


Figure 2: Item characteristic curves for an item adjusting for uncertainty in pseudo-guessing parameter estimates

Figure 2 shows an example of an item that has a difficulty parameter of approximately -1.8. Large differences can be seen in the ICCs, especially for ability levels less than +2. In addition to the uncertainty in the pseudo-guessing parameter estimates, there is also an increase in the uncertainty of the discrimination and difficulty estimates. This can be seen in Table 1 which compares the average standard errors of the parameter estimates for the 2PL and 3PL models for the 29 field test items that were shown in Figure 1. The largest impact is on the standard errors of the discrimination parameters, however even the location of the item is much more uncertain for the 3PL model.

Table 1: Average parameter estimates and standard errors for the 2PL and 3PL model with field test data

Model	Disc.	Disc. SE	Diff.	Diff. SE
2PL	0.559	0.052	0.124	0.139
3PL	0.929	0.158	0.541	0.181

Disc. = Discrimination, Diff = Difficulty,
PG = Pseudo-guessing, SE = Standard Error

The impact of 3PL IRT parameter estimates were discussed by Thissen and Wainer (1982). In this paper, the authors argue that the uncertainty in the pseudo-guessing parameters is not surprising given the amount of data found for low ability individuals, an idea exacerbated when dealing with field test data with less response strings. Their suggestion was to fit simpler models initially, and progress to more complicated models when the lack of fit was found to be inadequate (Thissen & Wainer, 1982). Also, a study by Kolen (1981) found that the pseudo-guessing parameters were not successfully estimated 92%, 53% and 39% across three different sets of items using the LOGIST program. Lastly, Lord (1980) suggests that the pseudo-guessing parameter should only be estimated if a significant lower tail in the ICCs are present in the range of ability levels found in the sample; this can be explored by looking at the empirical ICCs for an item. Again, when calibrating field test data, this problem would likely be amplified under small sample conditions.

The largest downside to fitting a 2PL model stems from the lack of fit compared to a 3PL model (CTB/McGraw Hill, 2008). However, the drastic increase in the uncertainty of parameter estimates under smaller sample size considerations is troubling, particularly when these parameter estimates are used directly when evaluating whether a field test item would be selected for an operational form. In addition, the uncertainty found with regard to the pseudo-guessing parameters influences the estimates and standard errors of the discrimination and difficulty parameters (Thissen & Wainer, 1982). This effect can be seen in Table 1 comparing the parameter estimates and standard errors of the discrimination and difficulty parameters for field test items. There is a need to evaluate alternative models to explore how well they are able to estimate item parameters that are less uncertain with comparable fit to the 3PL model, especially in small sample size conditions common in field test designs.

What is the guessing parameter?

From an achievement testing perspective, guessing can be conceptualized as when an individual does not definitively know the answer to a question, however attempts to answer the question anyway. There are likely many aspects that factor into the ability of an individual to arrive at the correct answer through guessing. The first and most straightforward would be through random guessing. This would be an act where the individual is unable to rule out any options and simply picks an answer at random. The act of random guessing in a standardized achievement test would likely occur when an individual does not have the ability to understand the question nor the answer options (Lord, 1980). This would result in the individual randomly selecting an answer out of all the alternatives.

Unfortunately, random guessing is likely not the answering strategy for many individuals. Instead, individuals likely attempt to make an informed guess by ruling out some of the answer options (De Ayala, 2013). For example, if an individual rules out two incorrect options out of four total response options, the individual would now have a 50% chance of answering the item correctly when answering randomly. Individuals in this group would likely be at a higher ability level, better understand the content of the question being asked to make the distractors less attractive, or some combination of the above.

The likely complex interaction of effects that influence guessing for an individual for a given item makes this term very difficult to rationalize and model. In addition, guessing would be best evaluated on individuals with very low ability levels, which may or may not be present in the given sample of individuals that the field test items are being tried out on. Lastly, with field test items, the distractors have commonly not been tried out to a large audience prior to the field test administration. This could result in some items with poor distractors, even for individuals with very low ability levels.

Of note, the idea of guessing only applies to situations where there are a finite and relatively small number of response options. The idea of guessing with open ended, constructed response, or in some cases technology enhanced items when there are an infinite or many response options is fundamentally different. With many more options to consider, the likelihood of an individual correctly guessing for a single item becomes smaller and across the many items on a single test becomes even smaller yet.

Alternative IRT Models

A simple alternative to the 3PL IRT model was explored by Han (2012). This article fixed the pseudo-guessing parameter to 1/[# of multiple choice (MC) options]. The simulation study identified numerous situations where estimates of the pseudo-guessing parameters were problematic. These areas occurred when the pseudo-guessing parameter was greater than 0.25 and with small sample size and longer test length (Han, 2012). These are all conditions that would likely be common when calibrating field test data. Han (2012) suggested there was little difference between the model fit with real data for the 3PL IRT model and the proposed method of fixing the pseudo-guessing parameter. In addition, this model has the advantage that the parameter estimates are comparable across items, not just for those with the same pseudo-guessing parameter estimates, a limitation of the varying pseudo-guessing parameter estimates (Lord, 1980).

However, for field test items, fixing the pseudo-guessing parameter to 1/MC options may not be flexible enough to account for small sample size conditions. Because field test items represent items being tried out for the first time, they may contain distractors that are less attractive to students, even those at low ability levels. It is argued that the phenomenon of "guessing" is likely very different in a field test environment. This may stem from unattractive distractors, ambiguity in the item directions or item stem, or even content not appropriate for a given grade level. These situations would represent instances that could not be adequately accounted for by a constant pseudo-guessing parameter as proposed

Table 2: Proportion of respondents selecting each option for a hypothetical multiple choice item

	A	B	C	D
All Respondents (AR)	.12	.68	.11	.09
Lowest 27% of Respondents (LR)	.18	.57	.13	.12

by Han (2012).

New Fixed IRT Models

We propose two slight variations of the fixed parameter 3PL IRT model that use empirical statistics from a distractor analysis to identify the fixed pseudo-guessing parameter value. More specifically, using the proportion of individuals that endorsed the least popular multiple choice option as an estimate of the lower bound of the item response function (i.e. the pseudo-guessing parameter) (Lindquist & Hoover, 2015). This lower bound will vary for each item and will automatically adjust if an item has a poor distractor. For example, if no individuals endorse one option, the pseudo-guessing parameter would be set to 0.

Two alternative ways to calculate the proportion of individuals endorsing the least popular multiple choice option are depicted in Table 2 for a single hypothetical item. One option will use all respondents (AR), as shown in the first row of Table 2, and the second option will use the lowest 27% of respondents (LR), as shown in the second row of Table 2. The value used to fix the pseudo-guessing parameter would be .09 and .12 for each option respectively. Sample size was the primary factor for exploring the two different approaches. With field test responses already being smaller compared to the number of responses for an operational form, further reducing the sample size to the lowest 27% of respondents could produce unstable proportion values. For this reason, the viability of using all respondents was also explored. Note, that the lowest 27% was used here as this data was already available to us. In practice, other percentages of the low ability individuals could be used, for example the lowest 25%.

The empirical method of fixing the pseudo-guessing parameter has many advantages over

Han's (2012) procedure of fixing the pseudo-guessing parameter to 1/MC options for field test data. First, it allows for the pseudo-guessing parameter to vary across items which may be desirable for items that may have an easy to rule out distractor. Secondly, the empirical data driven approach uses data already commonly calculated from a distractor analysis. Lastly, this method does lose the ability to directly compare items due to a lack of a constant pseudo-guessing parameter, however, with the 3PL model being popular for assessment data this is already a limitation of the common model.

Research Questions

The following research questions were explored:

1. To what extent is overall model fit affected by the four different model choices?
2. To what extent are estimates of the item parameters affected by the model choices?

Individual ability estimates were not explored here as they are commonly not of interest when calibrating field test items. The focus is specifically at the item level to evaluate which items are the best candidates for inclusion on an operational form. As a result, this is the focus of the current exploration.

Methodology

English Language Arts (ELA) and mathematics field test data were used from a single grade and a single administration year. The field test items were grouped together to create mini 'forms.' There were a total of 12 ELA field test items in a single mini-form and 30 mathematics field test items spread across two mini-forms. The two mathematics mini-forms were each taken by different students.

The field test 'forms' were administered after taking the state accountability exam. The field test mini-forms were randomized and spiraled within a classroom, meaning that students within a classroom or school did not all receive the same field test booklet. This will have the effect of eliminating, or at least minimizing, classroom or school effects attached to field test booklets through the random assignment like process.

In all calibration runs, the operational items were used as anchors to aid in estimation of the field test item parameters. In the case of the ELA field test items, there were 45 operational items that served as anchor items. Students with a valid scale score on the operational items were included in the analysis. The field test items were then appended to the end of the operational item string for those students who took both the ELA field test mini-form and the operational items. The structure described above is depicted in the matrix below.

$$\left(\begin{array}{cccc|cccc}
op_{1,1} & op_{1,2} & \cdots & op_{1,45} & ft_{1,1} & ft_{1,2} & \cdots & ft_{1,12} \\
\vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots \\
op_{p,1} & op_{p,2} & \cdots & op_{p,45} & ft_{p,1} & ft_{p,2} & \cdots & ft_{p,12} \\
 & & & & & & & \\
op_{p+1,1} & op_{p+1,2} & \cdots & op_{p+1,45} & & & & \\
\vdots & \vdots & \ddots & \vdots & & & & \\
op_{n,1} & op_{n,2} & \cdots & op_{n,45} & & & &
\end{array}\right)$$

A similar structure for mathematics calibration was performed. The main difference is that there are now two field test mini-forms appended to the 70 mathematics operational items. This structure is shown in the matrix below.

$$\left(\begin{array}{cccc|cccccccc}
op_{1,1} & op_{1,2} & \cdots & op_{1,70} & ft_{1,1} & ft_{1,2} & \cdots & ft_{1,15} & & & & \\
\vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots & & & & \\
op_{p,1} & op_{p,2} & \cdots & op_{p,70} & ft_{p,1} & ft_{p,2} & \cdots & ft_{p,15} & & & & \\
 & & & & & & & & & & & \\
op_{p+1,1} & op_{p+1,2} & \cdots & op_{p+1,70} & & & & & ft_{1,1} & ft_{1,2} & \cdots & ft_{1,15} \\
\vdots & \vdots & \ddots & \vdots & & & & & \vdots & \vdots & \ddots & \vdots \\
op_{p+q,1} & op_{p+q,2} & \cdots & op_{p+q,70} & & & & & ft_{q,1} & ft_{q,2} & \cdots & ft_{q,15} \\
 & & & & & & & & & & & \\
op_{p+q+1,1} & op_{p+q+1,2} & \cdots & op_{p+q+1,70} & & & & & & & & \\
\vdots & \vdots & \ddots & \vdots & & & & & & & & \\
op_{n,1} & op_{n,2} & \cdots & op_{n,70} & & & & & & & &
\end{array}\right)$$

The mathematics operational assessment had a sample size of about 7500 which served as an anchor to aid in estimation of the field test item parameters. The two field test booklets had sample sizes of 896 and 903. The ELA operational assessment also had a sample size of about 7500 and was used as an anchor to help estimation of the field test item parameters. The ELA field test booklet had a sample size of 898.

Models

Four models were fitted to the ELA and mathematics item responses. The baseline model considered was the 3PL model shown in equation (1) above. This served as the baseline model due to its wide use in practice. The alternative studied by Han (2012) was also fitted which fixed the pseudo-guessing parameter to 0.25 for all items (henceforth called FIX). This method was chosen to assess its viability for field test data for both mathematics and English language arts items. This is important as the distractors have not been as thoroughly tested for field test items; which may mean some options are not as attractive to the respondents.

Two new methods were explored that fixed the pseudo-guessing parameter to the proportion of individuals that endorsed the least popular multiple choice option. One method used all individuals who provided answers to the field test booklet to calculate the proportion (AR). The second option used the lowest 27% of individuals based on their scale score on the operational items (LR). Using the LR group would better reflect the group of respondents that would be most likely to guess on a given item. The proportion of respondents endorsing the least selected option would likely be larger for the LR group compared to the AR group. Sample size could be a concern when using the LR group, especially for field test data when sample sizes can already be small, resulting in a more unstable proportion.

The three models that fixed the pseudo-guessing parameter take the same general form as the 3PL model shown in equation (1), however c_j will not be estimated. Instead, it was fixed to the value specified for each model. For example, c_j was fixed to 0.25 for all items using Han (2012) methodology. For the methods fixing the pseudo-guessing parameter to the proportion of respondents endorsing the least popular option, the proportion was calculated from field test response data. Under these methods, the c_j parameter will still be allowed to vary from item to item, however, the estimation program will not estimate the parameter.

Analysis

The primary unit of analysis was the item level. This emphasis is due to the item level statistics being of most use when evaluating field test items for possible inclusion into an operational form. Parameter estimates and standard errors were used to assess degree of parameter estimate overlap for the competing models for each item. Confidence intervals will take the following form:

$$CI = est \pm 1.96 * SE \tag{2}$$

where est is the parameter estimate (either the a or b parameter estimates), and SE is the corresponding standard error. This will yield symmetric 95% confidence intervals around the parameter estimates. The 3PL model served as the baseline comparison due to the widespread use in practice.

Item characteristic curves (ICC) will also be used to highlight interesting and meaningful differences. These graphical displays represent the probability of an individual answering the item correctly given their ability level. Differences in these curves could have an impact on whether a field test item was selected to be included on an operational form. As mentioned before, consistent differences in an entire test form could result in different reported scores and impact the TCC.

Finally, aggregate standard errors and overall model fit were explored to highlight the model impact in general. AIC, which adjusts for the number of parameters being estimated, was used to measure model fit. AIC was calculated using the output from Parscale's log likelihood as follows (Akaike, 1974):

$$AIC = -2 * loglik + param * items \tag{3}$$

where *param* is the number of item parameters and *items* is the number of items.

Software

Calibration was done with Parscale, which uses marginal maximum likelihood for estimation, (Muraki & Bock, 1997) and verified with an R (R Core Team, 2015) IRT package, ltm (Rizopoulos, 2006). There were convergence problems with ltm in a handful of situations, as a result, the results for Parscale are reported. The results that did converge were comparable across Parscale and ltm. Data analysis was performed with R (R Core Team, 2015) and figures were created with the ggplot2 package (Wickham, 2009).

Limitations

Limitations for the current study stem from the real world data being used. As a result, the true parameter values for the discrimination, difficulty, and pseudo-guessing are not known and comparisons to the true values are not possible here. In addition, only mathematics and ELA tests were used in the current study. Care needs to be taken when attempting to use the study results outside of these two subject areas. Finally, a choice was made to study only the 3PL model and models that fix the pseudo-guessing parameter to a specified value. Alternative models such as the 2PL model, where the pseudo-guessing parameter is fixed to zero, or Rasch models offer additional alternatives. The choice to focus the study on 3PL and models that fix the pseudo-guessing parameter to a specified value was made to encourage better comparison between parameter estimates. In addition, the 3PL model is a widely used IRT model in practice for calibrating operational and field test items.

Results

Convergence

There were convergence problems when using the FIX method for the mathematics calibration. The model was unable to converge after 1000 cycles and 30 newton cycles. Default settings within Parscale were used and model convergence may have been achieved by altering these settings. The default settings were used throughout to achieve comparability across the models. These results are included below as parameter estimates were obtained, but caution in interpreting these results needs to be taken. The convergence problems empirically suggest that the FIX method is not flexible enough to accommodate the mathematics field test data. This idea will be explored in more detail below.

Parameter Estimation

Visual representations of the estimated field test item parameters using the four methods are shown below in Figure 5 and Figure 6, showing 95% confidence intervals (CI) for mathematics and ELA respectively. Exploring Figure 5 first, shows that on average the FIX method tends to have pseudo-guessing (PG) values slightly higher than the 3PL estimates and significantly higher than the AR or LR methods. This can also be seen in Table 3 where the average PG parameter was 0.226 for the 3PL model and less than 0.1 for the AR and LR methods.

The value of the PG parameter has direct impact on the estimates for the discrimination and difficulty parameters. As can be seen from Figure 5, the estimates for the discrimination and difficulty tend to be much smaller for the AR and LR methods with the width of the 95% CI much smaller. There are many confidence intervals that do not overlap when comparing the AR and LR methods to the other two. These significantly different parameter estimates would likely have a large impact on the ICC, particularly at the lower ability levels. Table 3 shows the same trend on the aggregate as well. For example, items tend to have smaller discrimination and difficulty estimates with much smaller standard errors for the AR and LR methods. The difference in the standard errors is particularly large for the discrimination parameter estimates where it is over twice as large for the 3PL model compared to the AR and LR methods.

ELA has many of the same trends as mathematics, but the effects are smaller. Figure 6c shows the PG parameter values and the difference between the 3PL estimates and the LR method were much smaller for ELA than mathematics. The FIX method tended to have higher PG values compared to the 3PL model which suggests that fixing the PG parameter to 0.25 in this case was empirically too large for the field test items. The impact of the larger PG values can be seen in the estimates of the discrimination and difficulty parameters. The FIX method tended to produce estimates that were larger compared to the other methods. Similar to the mathematics items, the AR and LR methods tended to produce slightly smaller estimates for the discrimination and difficulty parameters, however there was much more overlap in the 95% CI as shown in Figure 6 compared to mathematics.

Table 4 shows similar trends when looking at the aggregate of the items. On average, the PG values tended to be smallest for the AR and LR methods. However, compared to mathematics, the difference between the 3PL and LR methods was smaller. Also, the estimates for the discrimination and difficulty tended to be much more precise for the AR and LR methods compared to the other two methods. Although the FIX method had the smallest standard error for the difficulty parameter, the standard error of the discrimination parameter was larger than the AR or LR methods. This suggests that the estimation algorithm has more difficulty in estimating the discrimination parameter when the PG value is larger. Having the larger PG value also has the impact of forcing the discrimination parameter to be larger due to the restriction of range in the conditional probability of correctly answering the item.

Finally, Table 5 shows correlations between parameter estimates and standard errors for the mathematics (upper diagonal) and ELA (lower diagonal) for the field test items across methods (i.e. the methods were combined to calculate the correlations). Of note is the strong positive correlation between the parameter estimate and associated standard error of the discrimination parameter for both subjects. This suggests that larger parameter estimates are associated with larger standard errors for the discrimination parameter.

In addition, the strong positive correlation between the PG parameter estimate and the discrimination parameter and associated standard error is of interest. These correlations suggest that larger PG values are associated with larger discrimination estimates and standard errors. This further explains the trends shown in Figure 5 and Figure 6 and may suggest that simpler IRT models are warranted for field test data to increase precision and protect against overestimation of the PG parameters, a thought discussed by Thissen and Wainer (1982).

Table 3: Average IRT parameter estimates and their standard errors for the four competing models – Mathematics

Model	Disc.	SE Disc.	Diff.	SE Diff.	PG	SE PG
3PL	0.929	0.158	0.541	0.181	0.226	0.050
AR	0.575	0.060	0.156	0.131	0.039	0
FIX	1.037	0.151	0.567	0.126	0.25	0
LR	0.615	0.064	0.241	0.131	0.068	0

Note: Disc. = Discrimination, Diff = Difficulty,
PG = Pseudo-guessing, SE = Standard Error

Table 4: Average IRT parameter estimates and their standard errors for the four competing models – ELA

Model	Disc.	SE Disc.	Diff.	SE Diff.	PG	SE PG
3PL	0.954	0.131	-0.450	0.156	0.210	0.057
AR	0.785	0.074	-0.547	0.101	0.057	0
FIX	1.057	0.119	-0.193	0.096	0.25	0
LR	0.866	0.084	-0.442	0.101	0.115	0

Note: Disc. = Discrimination, Diff = Difficulty,
PG = Pseudo-guessing, SE = Standard Error

Table 5: Correlations between parameter estimates and the associated standard errors of the parameter estimates, upper diagonal are for mathematics and lower diagonal for ELA.

	Disc.	SE Disc.	Diff.	SE Diff.	PG	SE PG
Disc.	–	0.887	0.182	-0.236	0.559	0.083
SE Disc.	0.759	–	0.375	0.048	0.591	0.203
Diff.	0.167	0.354	–	0.411	0.176	-0.071
SE Diff.	-0.421	-0.033	-0.528	–	0.050	0.238
PG	0.476	0.721	0.333	0.096	–	0.450
SE PG	0.024	0.344	-0.205	0.682	0.316	–

Note: Disc. = Discrimination, Diff = Difficulty,
PG = Pseudo-guessing, SE = Standard Error

Item Characteristic Curves

Figure 3 presents an ICC for a single mathematics item where the 3PL and FIX methods give very little information on the low ability students due to the nearly flat ICC at low ability levels. The AR and LR methods show more discrimination across the ability scale, resulting in information available for a wider range of ability. For the ICCs with the four methods graphed (see Figure 3), the 3PL and FIX methods tend to exhibit similar patterns, while the AR and LR methods share similar patterns. The estimated difficulty parameters show a trend of decreasing across methods 3PL, FIX, LR, AR. These range from 1.03 for 3PL to .018 for AR, a difference of over 1 on the theta scale.

The ELA ICC shown in Figure 4 below demonstrates that the FIX method increased the PG estimate above that of the standard 3PL estimate. This particular item exhibits moderate differences along the ICC among the 4 methods. All alternate methods show similar discrimination and difficulty estimates for this item. The difficulty parameter for 3PL is -1.5 for all methods except FIX, which has a slightly higher value of -1.2. As shown earlier (see Tables 3 and 4) , the SE for the difficulty estimates are higher in the 3PL method than the other three. The SE for discrimination estimates are smallest for the AR and LR methods, resulting in greater confidence in the estimates.

Model Fit

AIC fit indices for the four alternative IRT model specifications are reported in Table 6. Smaller AIC fit indices indicate better model fit. The 3PL model had the best model fit of the four methods used for mathematics items (these fit indices include both operational and field test items in their calculation), while the LR method provided the best fit among the alternate estimation methods. Recall that the FIX method did not converge for the mathematics calibration run, which is likely contributing to the large AIC value. For ELA items, the 3PL model has the worst model fit and the LR method provides the best fit to the data. The AR and FIX methods are very similar.

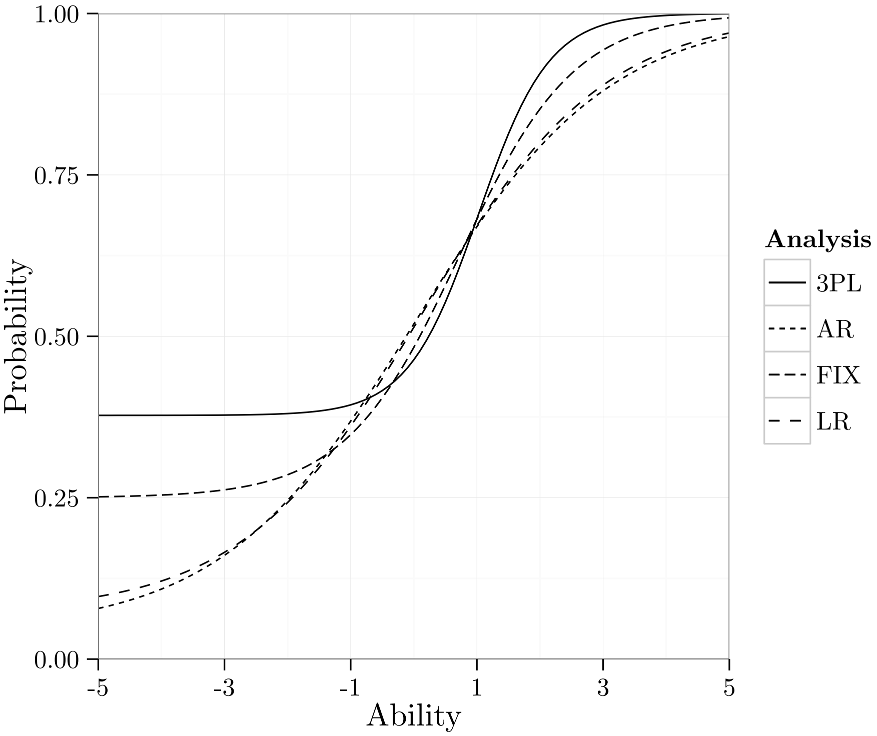


Figure 3: Item characteristic curves by analysis method – mathematics.

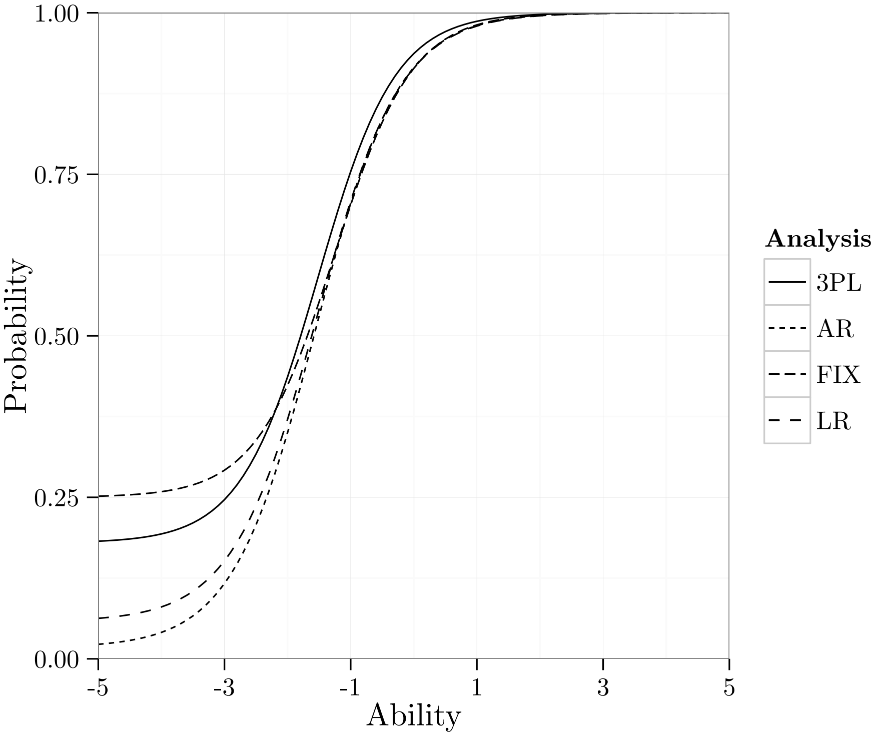


Figure 4: Item characteristic curves by analysis method – ELA.

Table 6: AIC values for the four competing IRT models

Subject	3PL	AR	FIX	LR
Mathematics	595947.5 (1)	596960.9 (3)	598217.5 (4)	596686.7 (2)
ELA	372997.9 (4)	368239.9 (3)	368239.3 (2)	368074.5 (1)

Note: Rankings within subjects are shown in parentheses.

These results suggest that the LR method is the most stable estimation method for these data, ranking second best for mathematics items and the best for ELA items. As identified by model fit, the FIX method does not seem flexible enough to adapt to field test data, and ranks toward the bottom in terms of fit. Lastly, the AR and LR method show similar fit, but AR was consistently worse. The LR method provides better fit and similar performance to that of AR, and may be superior if the N count can support stable distractor data.

Discussion

IRT models are data intensive procedures that require relatively large sample sizes for adequate and accurate estimation of item parameters. The data requirements further increase for more complicated IRT models such as the 3PL model which are widely used in practice (Baker, 1987; Hulin et al., 1982). In many situations, for operational assessments, the data limitations are likely not a problem. However, for field test items, the items are spread out over a pool of respondents which can put significant constraints on the sample size for a given field test item. These limitations on the sample size can have consequences on adequate and accurate estimates of the item parameters that are used for item selection for operational assessments, a thought explored by Thissen and Wainer (1982). The current study explored the effect of different 3PL IRT models on the estimation of IRT parameters. In addition to the traditional 3PL model, three alternatives that involved fixing the pseudo-guessing parameter to specified values were explored.

Study results showed that for field test items, there are large amounts of variation in the pseudo-guessing parameter estimates under a 3PL IRT modeling framework. The uncertain estimates for the pseudo-guessing parameters also increase the amount of uncertainty in the discrimination and difficulty item parameter estimates, a result discussed by Thissen and Wainer (1982). If parameter estimates are consistently over or under estimated, this could have severe implications for the overall test characteristic curve and may influence respondents score on the test. The correlations between the parameter estimates and the standard errors (see Table 5) highlights the association between larger pseudo-guessing parameters and the discrimination parameters and their standard errors. Having large pseudo-guessing estimates restricts the range of the item characteristic curve which has the effect of increasing the discrimination parameter. Exploring items with large discrimination parameters in Figure 5 and Figure 6 shows that these are cases when the methods diverge the most.

Of the three methods that fix the pseudo-guessing parameter to a specific value, calculating the proportion of respondents that endorsed the least popular multiple choice option with the lowest 27% of respondents performing the best. For this method, the estimates of the pseudo-guessing parameters tended to be much smaller compared to the 3PL and fixing to 0.25. However, the uncertainty in the discrimination and difficulty parameter estimates were also significantly smaller compared to the 3PL model. In addition, this model consistently provided good model fit compared to the other methods. For an individual item, the item characteristic curve provided a wider range of discrimination across the entire ability scale as well, which may be an added benefit.

The option to use the lowest 27% of respondents was specific to this study, but in practice a different theoretical value that represents the lowest ability on the given test could be used. For example, the lowest quarter or third of respondents could be used as an alternative. The benefit of using only the lowest ability students stems from the idea of “guessing.” Guessing theoretically would represent the situation where students are unable to rule out any of the distractors. This scenario would most likely to be true for the lowest ability students. In the current study, the fixed pseudo-guessing parameter was about twice as large for the lowest ability students compared to using all students.

Lastly, the method by Han (2012) which fixed the pseudo-guessing parameter to 0.25 (more generally to 1/# MC options) does not appear to be flexible enough to accommodate field test data. The poor fit indices, non-convergence for mathematics items, and the relatively small reduction in the standard errors for the discrimination parameter estimates suggest a method that does not adapt well to items that have not been fully tested.

Future Research

Future research in this area is warranted, particularly to explore different subject areas such as science, social studies, or written expression field test items. The difference in the behavior of these methods between the mathematics and ELA items suggest the concept of "guessing" may differ significantly between subjects.

Another area of research is how these model differences ultimately influence the reported scores, particularly for different reporting scales and when considering if field test item parameters were consistently over or under estimated. The large differences shown in the individual item characteristic curves in Figure 3 and Figure 4 suggest that if a consistent pattern in estimation for all or a large portion of items in an operational form, the overall probability of a student answering the item correctly could differ significantly. Relatedly, the different difficulty parameter estimates could influence which item is given to a respondent in a computer adaptive testing environment. Understanding how these different models can influence scores is an important practical consideration that needs further study.

References

Akaike, H. (1974). A new look at the statistical model identification. *Automatic Control, IEEE Transactions on*, *19*(6), 716–723.

Baker, F. B. (1987). Methodology review: item parameter estimation under the one-, two-, and three-parameter logistic models. *Applied Psychological Measurement*, *11*(2), 111–141.

Baker, F. B. & Kim, S.-H. (2004). *Item response theory: parameter estimation techniques*. CRC Press.

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. *Statistical theories of mental test scores*.

CTB/McGraw Hill. (2008, January). *Accuracy of test scores: why irt models matter*. Retrieved from https://www.ctb.com/ctb.com/control/openFileShowAction?mediaId=18747.0

De Ayala, R. J. (2013). *Theory and practice of item response theory*. Guilford Publications.

Downing, S. M. & Haladyna, T. M. (2006). *Handbook of test development*. Lawrence Erlbaum Associates Publishers.

Han, K. T. (2012). Fixing the c parameter in the three-parameter logistic model. *Practical Assessment, Research & Evaluation*, *17*(1), 1–24.

Hulin, C. L., Lissak, R. I., & Drasgow, F. (1982). Recovery of two-and three-parameter logistic item characteristic curves: a monte carlo study. *Applied psychological measurement*, *6*(3), 249–260.

Kolen, M. J. (1981). Comparison of traditional and item response theory methods for equating tests. *Journal of Educational Measurement*, *18*(1), 1–11.

Kolen, M. J. & Brennan, R. L. (2014). Test equating, scaling, and linking.

Lindquist, E. F. & Hoover, H. D. (2015). Some notes on corrections for guessing and related problems. *Educational Measurement: Issues and Practice*, *34*(2), 15–19. doi:10.1111/emip.12072

Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Routledge.

Mislevy, R. J. & Bock, R. D. (1990). *Bilog 3: item analysis and test scoring with binary logistic models*. Scientific Software International.

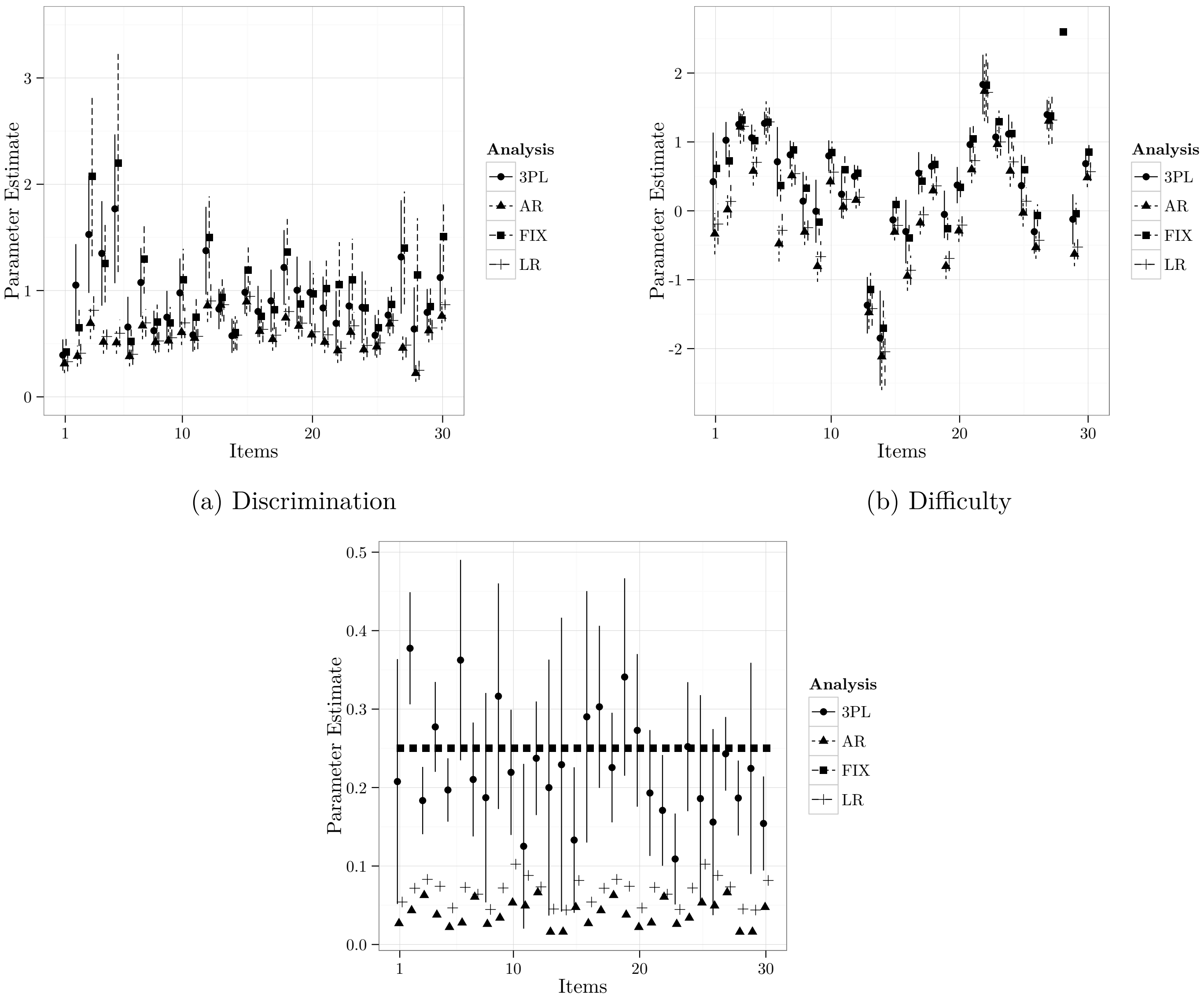
Muraki, E. & Bock, R. (1997). *PARSCALE: IRT item analysis and test scoring for rating-scale data*. Scientific Software International. Chicago, IL.

R Core Team. (2015). *R: a language and environment for statistical computing*. R Foundation for Statistical Computing. Vienna, Austria. Retrieved from https://www.R-project.org/

Rizopoulos, D. (2006). ltm: An R package for Latent Variable Modelling and Item Response Theory Analyses. *Journal of Statistical Software*, *17*(5), 1–25. Retrieved from http://www.jstatsoft.org/v17/i05/

Thissen, D. & Wainer, H. (1982). Some standard errors in item response theory. *Psychometrika*, *47*(4), 397–412.

Wickham, H. (2009). *Ggplot2: elegant graphics for data analysis*. Springer New York. Retrieved from http://had.co.nz/ggplot2/book

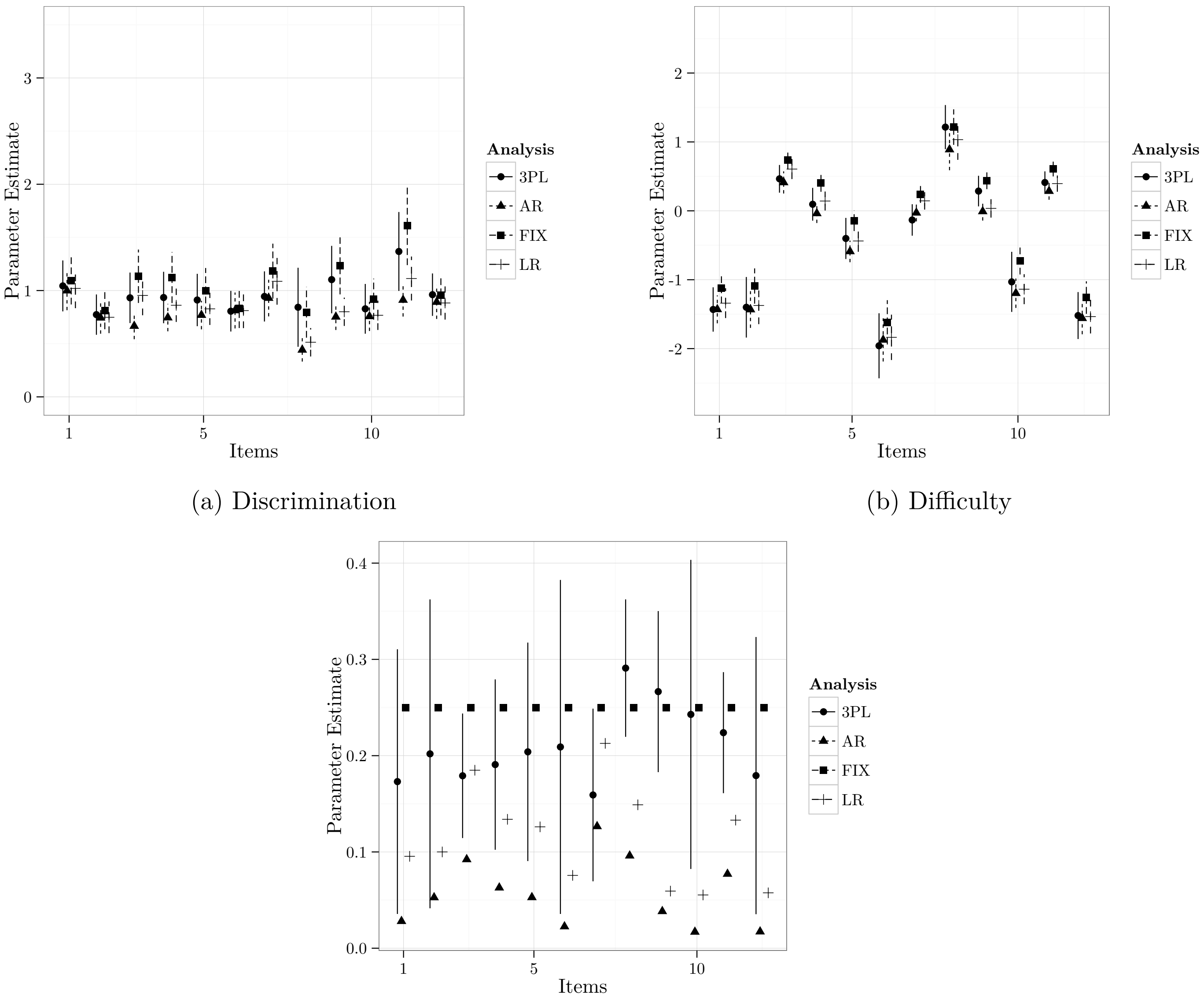


(a) Discrimination

(b) Difficulty

(c) Pseudo-guessing

Figure 5: Parameter estimates with 95% confidence interval – Mathematics



(a) Discrimination

(b) Difficulty

(c) Pseudo-guessing

Figure 6: Parameter estimates with 95% confidence interval – ELA