The University of Iowa

Iowa Testing Programs

Item Response Theory Estimation with Multidimensional Field Test Booklets
ITP Research Series

Brandon LeBeau
Wei Cheng Liu

ITP Research Series

2017.2

Item Response Theory Estimation with Multidimensional Field Test Booklets

Abstract

Field testing is a way for test developers to gather data on the performance of items prior to appearing on an operational assessment. One difficulty with field test data is that many items end at this stage and do not make it to an operational assessment, which can have the effect of reducing the sample size for these items. Smaller sample sizes can make it increasingly difficult to achieve useful and accurate estimates from an item response theory model. Grouping items of similar content, for example written expression and vocabulary items, into a single field test booklet can help to maximize the number of respondents for a given item. However, this can lead to multidimensional booklets that can provide complications for common item response theory methods in addition to the concern over small sample sizes. This case is explored with real field test data to attempt to understand the impact of multidimensionality on item calibration in small samples. Implications for test developers and psychometricians will be discussed.

In the late 1990s, Downing and Haladyna (1997) contended that most studies and discussions surrounding validity were conducted after the tests were administered and there had been a lack of attention given to validity evidence associated with development of test items. This notion of validity does not need to solely concern the content of the test items, but also should apply to the item analyses as well. This is particularly important when item response theory is used as a part of the item analysis of newly tried out items for at least two reasons. First, these items represent those that have not appeared on an operational form, therefore may not be final with regard to their content delivery, distractor validation, or content appropriateness. Secondly, estimation of the item parameters under an item response theory framework is asymptotic, requiring large samples which may not be possible from a field test design.

From this perspective, test developers may need to be creative to ensure adequate sample sizes while still trying enough new items to fulfill their content requirements. One strategy that may be useful to test developers is to include two related, but different constructs within a single item tryout. However, although this may be attractive from a test developer framework, there are now a few options to analyze these new items using item response theory. For example, the items could be split by the constructs and analyzed separately or they could be combined and analyzed together in a single calibration. If the items are combined, this may violate the unidimensional assumption common for traditional item response theory models.

This study explores this issue in more detail with respect to real field test data and the impact of multidimensional tryout items on the estimated item parameters from an item response theory analysis. In the current study, written expression and vocabulary items

were administered in a single field test and would likely represent overlapping but somewhat distinct constructs. Various methods of analyzing this data structure are explored, including methods that model the multidimensional nature of the data and others that assume the data are unidimensional. This extends current literature in that it represents small sample conditions with real field test data. Much of the current item response theory literature explores data from operational forms and the analysis of field test data present new challenges when applying item response theory models.

Field Testing of New Items

In large scale testing companies, new items are consistently needed to replace overexposed items or for development of new forms. Field testing of these new items is a common practice to collect data for item analysis to determine the performance of these items (Livingston, Downing, & Haladyna, 2006). Many standardized testing agencies employ item writers, who are content experts, to write new items that match content standards to be evaluated on an assessment. As can be imagined, there are many more items written than actually make it onto an operational test form (Downing & Haladyna, 2006). This has a few implications, first, that many more items than are needed are tried out through field testing, and secondly, these items are spread out over a fixed set of individuals that may result in a smaller number of responses for each new item that is field tested.

There are many ways to gather data on field test items. One way is to embed items directly in the middle of the operational items. 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 items 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 to create 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 test. This has the advantage of not fatiguing individuals when taking the operational test, however has the disadvantage of the individuals now knowing these are field test items (or that these items no longer influence their score). This may impact the response strings due to individuals not giving 100% effort (Downing & Haladyna, 2006), which may further reduce small sample sizes 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 arise.

Limitations of Field Test Designs

The primary objective when conducting field tests is to gather item statistics to make informed decisions regarding whether the item(s) 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 item response theory (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, particularly due to the asymptotic estimation methods that rely on large sample estimation theory (De Ayala, 2013; Lord, 1980). This is especially true when a 3PL model is used for estimating the IRT item parameters (a process called calibration). The strengths and weaknesses of this model will be presented below for field test data (or more generally for small samples) with particular attention to the unidimensionality assumption underlying the IRT model.

IRT methods for equating or linking.

Item Response Theory (IRT)

Data collected from field tests are analyzed for purpose of making informed decisions regarding the suitability of newly written items on operational forms. These analyses provide three kinds of information about the items - difficulty, discrimination and differential item functioning (Livingston et al., 2006). Other than relevant statistics from classical statistical theories, like proportion of respondents answering the item correctly and biserial correlations, item parameter estimates from IRT are also utilized for such decisions. In fact, IRT has become the new standard in item analysis due to its advantages over classical test theories, for example, when the same ability scale is used, invariance of item parameters across groups of respondents and invariance of respondent's ability across tests of the same underlying construct (De Ayala, 2013; Lord, 1980). IRT offers greater model flexibility and the ability to estimate item parameters and respondent's ability at the same time. Apart from item analysis, IRT has also be extensively applied to equating and linking of different tests. Kolen and Brennan (2004) provides a thorough discussion on using IRT methods for equating or linking. These advantages are afforded through the application of strong assumptions, for example, unidimensionality which is the focus of this study. Therefore, understanding how IRT models behave when assumptions have not been met is an important topic for psychometricians and applied researchers alike.

A commonly used IRT model is Birnbaum's three parameter logistic model (Birnbaum, 1968):

$$p(x_{ij} = 1|\theta_i, a_j, b_j, c_j) = c_j + (1 - c_j)\frac{1}{1 + e^{a_j(\theta_i - b_j)}}. \quad (1)$$

The equation above models the likelihood of a respondent correctly answering item j given the respondent i's' ability (θ_i), the item's difficulty (b_j), the item's discrimination (a_j), and the pseudo-guessing parameter of the item (c_j). These three item parameters make up the

item response function which is a mathematical representation of the likelihood of respondent answers a given item correctly given their ability. See De Ayala (2013) and Lord (1980) for a more in depth introduction to IRT.

IRT Assumptions

A common assumption of the three parameter IRT model shown above is unidimensionality. This assumption states that all items calibrated together all underlie the same construct or latent person variable (De Ayala, 2013; McDonald, 2013; Reckase, 2009). As an example, the items on a mathematics achievement test would be assumed to all belong to a single latent variable that could be adequately accounted for by an individual's mathematics achievement. If the response pattern for the items cannot be assumed to follow a single latent person variable (i.e. mathematics achievement and calculation speed), the unidimensionality assumption underlying the IRT model above would likely be violated.

Research exploring the implications of violating the unidimensionality assumption has shown that the model is relatively robust (Harrison, 1986; Henning, Hudson, & Turner, 1985) to small violations. However, others have stated that increasing the number of multidimensional items can lead to complications (Oshima & Miller, 1992) and others have shown as the dominance of the general factor decreases, the root mean square differences increase for the discrimination and difficulty parameters (Drasgow & Parsons, 1983). In addition, a study by Reckase (1979) showed that the size of the first eigenvalue is directly related to the 3PL discrimination estimates, variability of the 3PL difficulty estimates, the 1PL probability of fit and mean square deviations for 1PL and 3PL models. Other researchers have stated that achievement test data will not meet the unidimensionality assumption (Ackerman, 1994). A rule of thumb by Reckase (1979) was that "for acceptable calibration, the first factor should account for at least 20 percent of the test variance." Therefore, careful evaluation of this assumption is needed to ensure adequate and useful parameter estimates.

Another assumption found with IRT is the idea of independent responses called the local independence assumption (Reckase, 2009) or conditional independence (De Ayala, 2013). This assumption assumes that the response by an individual on a single item does not influence their responses to other items. The term local does not refer to each individual, but rather refers to groups of individuals with the same ability level. For example, all individuals with an ability of +2 are assumed to have answered independently, but their responses may be correlated with the responses of individuals with different ability levels. Another way to think of this assumption is that the correlation between item scores is due to variation in the ability of individuals, not due to the items themselves (De Ayala, 2013; Reckase, 2009).

It can sometimes be thought that the unidimensionality and independence assumptions are directly related. For example, if independence is not tenable, it is common to assume the correlation in item scores at a given ability level is due to the presence of a second (or more) constructs that are not adequately accounted for. For instance, when mathematics ability is not the only construct being measured, but also English comprehension or the ability to do computations quickly (referred to as speeded tests) are also needed to answer the items correctly. This second construct can produce a dependency in the item scores within an ability level. However, as De Ayala (2013) describe, other explanations are possible. A set of

interrelated or hierarchical items in which answers depend on correctly answering previous questions could be one such alternative explanation (De Ayala, 2013).

Detecting Multidimensionality

There have been many methods for detecting multidimensionality throughout the literature. Exploratory and confirmatory factor analysis have been used extensively in the exploration of the dimensionality of an assessment (Green & Yang, 2015; Reckase, 1979). Other methods compare relationships between single pairs of items (Chen & Thissen, 1997) and others, such as DIMTEST tests for dimensionality by considering all items on the test (Stout, 1987). Even others use the eigenvalues as a guide to the dimensionality question (Reckase, 1979, 2009). A study by Finch and Habing (2007) offers some suggestions for when DIMTEST or normal ogive harmonic analysis robust method (NOHARM) should be used and Green and Yang (2015) argue for the use of factor analytic methods to determine dimensionality.

Multidimensional IRT

When there is evidence that the unidimensionality IRT assumption has not been met, there is evidence that more than one ability distribution is present for each individual representing more than one construct. These multiple constructs can be modeled directly through the use of multidimensional IRT (MIRT). MIRT is an extension from the 3PL model given in Equation (1) shown above. The MIRT 3PL model (M3PL) takes the following form (Reckase, 2009):

$$p(x_{ij} = 1|\boldsymbol{\theta_i}, \boldsymbol{a_j}, d_j, c_j) = c_j + (1 - c_j)\frac{1}{1 + e^{\boldsymbol{a_j}\boldsymbol{\theta_i^T} + d_j}}. \tag{2}$$

In the M3PL model, much is the same from the unidimensional 3PL (U3PL) model shown in (1). The extension to multidimensional space increases the dimensions of the discrimination and ability for each individual, these are shown in Equation (2) as $\boldsymbol{a_j}$ and $\boldsymbol{\theta_i}$ respectively. These vectors would now be 1 x m where m is the number of dimensions in the coordinate space. Lastly, the new term, d_j, is related to the difficulty parameter from the U3PL model, however now represents $-ab$ found when expanding the numerator of the U3PL model shown in Equation (1); $a(\theta - b) = a\theta - ab$. However, due to having more than one ability dimension, d_j is commonly transformed by (Reckase, 2009)

$$b = \frac{-d}{\sqrt{\boldsymbol{a}\boldsymbol{a^T}}} \tag{3}$$

The pseudo-guessing parameter, c_j, is equivalent in both models and only one is commonly estimated with the M3PL model.

The MIRT model shown above in Equation (2) can be specified in ways that mimic exploratory and confirmatory factor analysis (Chalmers, 2012). The confirmatory MIRT approach would be useful when a priori the researcher knew which constructs were present for each item. This information could be obtained from the test blueprint, comparing the item to specific standards, or even consulting with experts. In cases when a confirmatory MIRT approach is used, the $\boldsymbol{a_j}$ and $\boldsymbol{\theta_i}$ vectors for each individual may be reduced as some components would be fixed to zero when it is identified that a specific factor is unrelated to that item.

Intersection of Multidimensionality and Field Testing

To date, no research has specifically explored the intersection of multidimensionality and field testing. This situation would be relatively unique in that small sample sizes combined with multidimensional field test data could provide additional problems over and above what has been traditionally explored in the IRT literature. The Monte Carlo literature on violating the unidimensionality assumption have commonly had sample sizes around 1,000 and the number of items have ranged from 30 to 70 (Drasgow & Parsons, 1983; Harrison, 1986). In addition, the items loading on the common factors ranged anywhere from 0.4 to 0.8 assuming a simple structure (i.e. that the items load solely on a single common factor), but the common factors were allowed to be correlated (Drasgow & Parsons, 1983; Harrison, 1986). These design considerations may not be similar for field test data. With field test data, sample sizes may be smaller, there may be many more items to calibrate, especially if the operational items are included in the calibration, and the correlation between the field test items and the common factors may be smaller than has been studied in the simulation literature. Therefore, these constraints may put more strain on the unidimensionality assumption and should be explored empirically.

Another common design consideration in the simulation literature surrounding multidimensionality is the number of common factors assumed to underlie the data. These have been set to range between four and eight (Drasgow & Parsons, 1983; Harrison, 1986), but additional study of the number of common factors, particularly with smaller numbers may be useful. Correlations between the common factors and the general factor were also manipulated in these two studies (Drasgow & Parsons, 1983; Harrison, 1986). Even though these studies do include situations where the common factors and general factor are perfectly correlated, a condition where the unidimensionality assumption would be met, only small deviations were made in the design in which the correlation were consistently larger or smaller. There were no variation in these correlation structures. This again may represent conditions that differ when trying out new items through a field test.

Finally, a study by Henning et al. (1985) explored the unidimensionality IRT assumption of a language assessment, the English as a Second Language Placement Examination (ESLPE). This assessment was comprised of 150 items from 5 subtests including: listening comprehension, reading comprehension, grammar accuracy, vocabulary recognition, and writing error detection. These subtests would likely all be related, but somewhat distinct in their underlying constructs and would mimic the current data from this study in some regards. In addition, the correlations between the subtests ranged from around 0.60 to 0.90. Henning et al. (1985) used a Rasch IRT model to these data and found that the model was robust to the unidimensional IRT assumption and concluded that this model may be appropriate for Language assessments in general.

Research Problem

This study explores the implications of calibrating multidimensional data. More specifically, this study extends the current literature with the focus on field test booklets. The data available from field test booklets tends to be much smaller and more sparse compared to data from operational items. As such, the multidimensionality problem may be exacerbated

in small sample situations. In addition, it is commonly of interest to use a single field test booklet to test as many items as possible. As such, stand-alone items can be a good way to fill out a field test booklet to maximize the amount of information gained from the single booklet. Including items that violate the unidimensionality IRT assumption may be needed to achieve this.

The following research questions will be explored:

1. To what extent do the field test booklets violate the unidimensionality assumption?
2. To what extent do violations to the unidimensionality assumption affect parameter estimates of the field test items?
3. To what extent do variations in the operational items used for anchoring affect research question 2 above?

Methodology

Data for this study are part of a larger field test administration. The field test administration was given to students after taking a large scale standardized achievement test, referred to as a field test booklet. The field test booklets used in the current study were administered to a single grade and were comprised of both written expression and vocabulary items. The field test booklets were spiraled amongst students in all participating school districts to help control for classroom effects. The spiraled design means that students within a classroom did not all receive the same field test booklet. The majority of the items on the two field test booklets were written expression (70%) compared to vocabulary items (30%).

The structure of the field test booklets was initially explored using confirmatory factor analysis (CFA) (Kline, 2011). CFA was used instead of exploratory factor analysis as it was known a priori which items were written expression versus vocabulary items based on the item development process and the test blueprint. Three competing models were explored for each of the booklets, a single factor model, a two factor model, and a bi-factor model that includes a general factor. For a more thorough discussion of the bi-factor model, see DeMars (2013) and Holzinger and Swineford (1937).

IRT Models

Three separate calibrations using three parameter logistic IRT models (3PL) were performed on the data (De Ayala, 2013). The field test items were calibrated with different operational items used as an anchor, including written expression, vocabulary, and mathematics operational items. Written expression and vocabulary operational items were selected as these two item types made up the items in the field test booklets. The mathematics operational items were selected as a more extreme case of the violation to the unidimensionality assumption.

Calibrations were also done with just written expression or vocabulary items. For example, instead of calibrating all the field test items, only the 70% of field test items that were written expression were used. These calibration methods served as the comparison group because these would represent the conditions that best satisfy the unidimensionality IRT assumption. The other calibrations methods would have some component of multidimensionality to them which may influence parameter estimates.

Table 1: Summary information for the anchor items used in the calibration of the field test items.

Subject	Number of Items	Sample Size
Mathematics	65	18,064
Vocabulary	39	10,284
Written Expression	43	7,942

In addition, 2PL models were also explored to examine the impact of model choice on the estimation. The smaller sample sizes corresponding to field test data can create greater uncertainty in the parameter estimates for a 3PL model (Lord, 1968; Hambleton, Jones, & Rogers, 1993; Jones, Smith, & Talley, 2006), particularly the pseudo-guessing parameter (Thissen & Wainer, 1982). Therefore, the simpler 2PL model was fitted which fixed the pseudo-guessing parameter to zero for all items.

The data setup for the calibrations were as follows:

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

where op_* represent the operational items and ft_* represent the field test items. The sample size for the two field test booklets were 1000 and 1003. The sample size and number of items for the group of anchor items varied as a function of the subject area. These can be seen in Table 1. The largest sample sizes and number of items were for the mathematics operational test and written expression operational items had the smallest sample size. De Ayala (2013) suggests that sample sizes of at least 1000 are needed for proper estimation using the 3PL IRT model and at least 500 are needed for the 2PL model. Both the operational items and field test items had sample sizes that meet these requirements.

Lastly, M3PL models were also explored and compared to the U3PL models. Three different confirmatory MIRT models were run, one for each of the three sets of base items considered above, mathematics, written expression, and vocabulary. The MIRT model using mathematics operational items was specified as having three constructs, one each for the mathematics, written expression, and vocabulary items. In contrast, the MIRT models using written expression and vocabulary operational items only had two constructs. The U3PL models were refitted with the new software to ensure estimates were in the same metric and to avoid differences in software implementation. More detail on the software used for estimation is presented below. A total of six new models were fitted and the parameter

estimates will be compared. M2PL and U2PL models were also fitted for comparison as well.

Analysis

Outcomes of interest included pairwise comparisons of the estimates and standard errors of the field test items for each calibration run. It is expected that based on the CFA results, the estimates for the FT items when calibrated with WE and vocabulary items will be more similar than when calibrated with mathematics items. Item characteristic curves (ICCs) were also used to show differences in the probability of a respondent answering a given item correctly based on their ability level. This is important as different item parameter estimates can yield similar ICCs.

To better quantify differences in ICCs, the unsigned area difference between ICCs will be used following procedures first developed for differential item functioning (DIF) by (Raju, 1988). In this method, two ICCs that diverge would have a greater unsigned area difference compared to those that are similar would have an unsigned area difference near zero. As shown by (Raju, 1988), the exact unsigned area difference for U3PL models that have different pseudo-guessing parameters would be infinite, therefore to overcome this limitation, (Kim & Cohen, 1991) depicted the unsigned area difference over closed intervals. The current analysis calculated unsigned area difference between two ICCs on the ability scale between -5 and 5 over 101 equal intervals. The comparison ICC was the written expression only and vocabulary only calibrations for the written expression and vocabulary field test items respectively. These were chosen as the comparison due to these best representing the unidimensional assumption. For more details on the method, see (Kim & Cohen, 1991) and (Raju, 1988).

Software

Calibration was done with Bilog-MG (Mislevy & Bock, 1990) for the unidimensional IRT models and an R (R Core Team, 2015) package, mirt (Chalmers, 2012), was used for the MIRT models. The one exception to this was that unidimensional IRT models for the comparison to the MIRT models were run with mirt. Data analysis was performed with R (R Core Team, 2015) and figures were created with the ggplot2 R package (Wickham, 2009).

Limitations

The data used for these comparisons are empirical data. Therefore, the true parameter values are unknown and the true model is not known. However, with this in mind, it is argued that there are calibration options that would yield unidimensional constructs assumed by the IRT models discussed, particularly the data conditions where the written expression or vocabulary field test items are isolated. These are treated as the comparison group throughout.

Table 2: Fit indices for three models fitted to each booklet.

Booklet	Model	χ^2	RMSEA	CFI
	One Factor	228.863 (119)	0.030	0.962
1	Two Factor	198.639 (118)	0.026	0.972
	Bi-factor	150.774 (102)	0.022	0.998
	One Factor	204.736 (119)	0.027	0.977
2	Two Factor	178.517 (118)	0.023	0.984
	Bi-factor	139.035 (102)	0.079	0.990

Note: RMSEA = root mean square error of approximation, CFI = comparative fit index

Results

Before performing the CFA analyses, the eigenvalues were first explored for each field test booklet. For both booklets, the largest eigenvalue accounted for just over 20% of the variation in the booklet. The first booklet had four eigenvalues greater than 1 and accounted for about 40% of the total variation and the second booklet had 6 eigenvalues greater than 1 and these accounted for just over 50% of the total variation. These eigenvalues provide some evidence of unidimensionality concerns, however, it also shows that the first eigenvalue is representing a large proportion of the total variation and fits with Reckase's (1979) suggestions.

Table 2 shows the fit statistics of the three competing CFA models for the two field test booklets. For both field test booklets, the model fit statistics indicated that all three CFA models fit the data well. While the fit statistics and chi-square difference test indicated the bi-factor model fit the data best over the other two models, the increased complexity of the bi-factor model was not justified by the modest improvement in the RMSEA. Furthermore, the bi-factor model and its corresponding factor loadings of items did not align to the conceptualization of the written expression (labeled as WE) and vocabulary (labeled as V) assessment. The two factor model was used for the rest of the analysis in this study.

For the two factor model, the correlation between written expression and vocabulary factors were 0.85 and 0.84 for booklet one and two respectively. This provides evidence that these two domains were similar with considerable overlapping information ($0.85^2 = 0.72$ or 72% of variation are similar). These values were also larger than the correlation between operational forms of written expression and vocabulary assessments. The factor loadings for each item can be seen in Table 3. This table shows the factor loadings of the V items were higher than the loadings for WE items for both of the field test booklets. Similar loading patterns were observed in operational forms as well (not shown). This may be due to the nature of the V items having a simpler structure compared to the WE items, for example, a V item could be one asking what the definition of a specific word whereas a WE item may be attempting to make connections within a text.

The table also shows that two factor loadings for WE items were not significant, item 6 and item 5 for booklet one and two respectively. When calibrating items, these two WE items were excluded. Convergence problems were found in a handful of the calibration runs

Table 3: Confirmatory factor loadings for the two field test booklets with standard errors.

		Booklet 1		Booklet 2	
	Item Number	Estimate	SE	Estimate	SE
Written Expression	1	0.514	0.042	0.400	0.046
	2	0.469	0.045	0.309	0.047
	3	0.336	0.058	0.386	0.048
	4	0.178	0.046	0.186	0.051
	5	0.558	0.040	0.054	0.053
	6	-0.012	0.072	0.741	0.040
	7	0.638	0.045	0.204	0.054
	8	0.725	0.037	0.317	0.053
	9	0.337	0.045	0.276	0.048
	10	0.500	0.040	0.350	0.046
	11	0.356	0.044	0.506	0.046
	12	0.624	0.042	0.474	0.046
Vocabulary	13	0.546	0.040	0.870	0.024
	14	0.616	0.039	0.846	0.025
	15	0.842	0.032	0.789	0.029
	16	0.755	0.036	0.788	0.029
	17	0.679	0.037	0.750	0.030

Note: SE = standard error

with default settings in Bilog-MG. The addition of a prior distribution for the difficulty parameters and adding ridge constants alleviated convergence problems. For comparability, similar settings were applied to the other calibrations.

Three Parameter Models

Written Expression

Figure 1 shows boxplots of the parameter estimate differences between different groups of operational items used as base items for written expression field test items. The comparison group for each figure is calibrating WE operational items with WE only field test items, which should represent the condition closest to satisfying the unidimensional assumption. As a result positive values represent estimates larger than the WE only group and negative values represent estimates smaller than the WE only group and values of 0 represent the same estimate. The median value for the differences in parameter estimates for the discrimination and difficulty parameters were all very close to zero regardless of the operational items used in calibration. The biggest implication for the discrimination and difficulty parameters is the increased level of variation in the parameter estimate differences, particularly when using mathematics items for calibration, which resulted in the worst fit out of all the base items. There was also increased variation in parameter estimate differences when vocabulary items were used as the base items.

The pseudo-guessing parameter estimates had much larger parameter estimate differences, again particularly when using mathematics items as base items for calibration. In this case, the parameter estimates were significantly overestimated. This may have been the reason for the increased variation and some large parameter differences for the discrimination and difficulty parameter estimates as well. Using vocabulary and written expression items as base items yielded median estimates similar to when WE only items were calibrated.

Figure 3 shows 95% confidence intervals for each of the 22 written expression field test items calibrated based on which calibration group was used. This figure gives a more detailed version of what was shown in Figure 1. For example, you can see that all of the pseudo-guessing parameters are estimated much smaller when using mathematics items as the base calibration group. Differences are also found for the discrimination and difficulty parameter estimates when using the mathematics items as the base calibration group. On average, the other three calibration groups have similar estimates for all three parameters, which is also shown in the boxplots from Figure 1.

Lastly, different IRT parameter estimates can yield similar item characteristic curves and these are shown for two WE items (items 10 and 22 respectively) in Figure 5. Item 10 in Figure 5, represents an item that is relatively similar between the four methods and very similar when the ICC for mathematics items is excluded. The WE and WE Only groups are virtually indistinguishable for this item. The ICC for the V group tends to be a bit flatter compared to the WE curves and approaches the mathematics curve for high ability levels. The strong difference between the pseudo-guessing parameters is also evident from the ICCs. The second item shown in Figure 5 shows a similar trend as the first item, but the differences between the WE curves and the V and mathematics curves is more pronounced. The larger pseudo-guessing parameters for the V, WE, and WE Only curves has the effect

of increasing the discrimination parameter estimate for this item.

Descriptive statistics for the average difference in the unsigned area between ICCs can be seen in Table 4 where the written expression only was used as the comparison ICCs. From the table, one can see that on average the ICCs when calibrating the field test items with mathematics base items have the most difference in the area with reference to the written expression only calibrations. Not surprisingly, the smallest area was when the calibration again using written expression with the vocabulary field test items. Using vocabulary base items for calibration also produced ICCs that on average had small deviations from the written expression only group.

Table 4: Descriptive statistics of unsigned area between item characteristic curve for different calibration methods.

Subject	Comparison Group	Mean	SD	Min	Max	Median
Written Expression	Mathematics	1.11	0.40	0.27	1.68	1.19
	Vocabulary	0.23	0.14	0.05	0.56	0.22
	WE with V	0.12	0.13	0.01	0.43	0.07
Vocabulary	Mathematics	1.55	0.37	0.91	1.99	1.58
	Vocabulary	0.17	0.17	0.01	0.54	0.06
	V with WE	0.41	0.15	0.18	0.66	0.44

Note: WE = written expression, V = vocabulary, SD = standard deviation, Min = minimum, Max = maximum

Vocabulary

Figure 2 shows the differences between parameter estimates of the vocabulary field test items using mathematics, vocabulary, and written expression operational items compared to calibrating the items with vocabulary operational and field test items only. The differences in parameter estimates when mathematics operational items were used are significant, particularly for the discrimination and difficulty parameters. In general, these parameters were larger compared to the calibration with vocabulary only items. Not surprisingly, only small differences were found between the vocabulary only and vocabulary calibrations. The exception to this is in regard to the pseudo-guessing parameter estimates which tended to be smaller for the vocabulary calibrations. When using written expression base items, the differences in parameter estimates were slightly larger compared to those with vocabulary base items, but on average were consistent.

Figure 4 depicts 95% confidence intervals of the parameter estimates for the ten vocabulary field test items by the four calibration groups. From this figure, the larger parameter estimates for the discrimination and difficulty parameters is apparent. The large confidence intervals for the pseudo-guessing parameters is also of some concern as this can have influence on the uncertainty and estimate of the other two parameter estimates. Also from the figure, the V and V only parameter estimates tend to cluster together for most of the ten items across the three parameter estimates. The largest differences between the V and

V only estimates tend to be with the pseudo-guessing parameters, although the confidence intervals do tend to have significant overlap.

ICCs are shown in Figure 6 for two vocabulary field test items (item five and nine) by the different calibration groups. Similar to the written expression field test items, there are significant differences in the ICCs when mathematics items were used as the base group compared to the other three methods. These differences are particularly strong at the lower ability levels for these two items. Item nine, has similar ICCs for the V, V only, and WE calibration groups and would likely not represent significant concerns over the probability of a correct response over the ability metric. Item five does show differences between the V only group and the V and WE groups, particularly for higher ability levels. Interestingly, the ICC mimics the V and WE curves below an ability of zero and when ability is greater than zero mimics the mathematics curve.

Table 4 shows descriptive statistics for the area between ICCs using the vocabulary only group as the comparison. Similar to the written expression analysis, the mathematics operational items had the greatest average distance between the ICCs and this difference was even more severe compared to the written expression items. In addition, the minimum value for the mathematics comparison group was larger than the maximum value for the other two groups, suggesting strong deviations. The other two were much closer to the vocabulary only ICCs with vocabulary base items being the closest on average.

Two Parameter Models

The results for two parameter IRT models were similar to that of the 3PL (results not shown). For the 2PL model, the parameter estimates for the written expression field test items using mathematics base items tended to be different compared to the WE only group. The discrimination and difficulty parameter estimates using mathematics base items tended to be larger with the 2PL model compared to the 3PL model. However, with no estimation of the pseudo-guessing parameter, the ICCs tended to be more similar across the different calibration groups.

Similar trends were found when exploring the unsigned area difference between ICCs for the 2PL model compared to the 3PL model shown in Table 4. More specifically, calibrations with mathematics base items performed the worst and even the minimum average unsigned area difference for mathematics items was larger than the maximum for the other two comparison groups. Secondly, the calibration most similar to the comparison group was the base group of items that matched the field test subject matter. For example, when calibrating WE items, the calibrations that used the WE base items were more similar than using vocabulary items.

Multidimensional Models

Figure 7 and Figure 8 show the parameter estimates for the six different calibrations, three using M3PL and three using U3PL models for written expression and vocabulary field test items respectively. Exploring the written expression field test items first (Figure 7) shows that there is significant variation in the parameter estimates across the six calibration methods, particularly for the pseudo-guessing parameter and items in field test booklet two (i.e.

items 14 to 22) for the discrimination parameter. The difficulty parameter did show more consistency in estimation for many of the items. The variability in the discrimination parameter for items 14 to 22 was similar to the increased variability for these items shown in Figure 3. In addition, the estimates tended to be larger for the M3PL models compared to the U3PL models for almost all items and across parameters, particularly when the comparison is focused within subject groups (e.g. mathematics operational items).

The M3PL and U3PL calibrations with written expression items produced very similar estimates for the three parameters as shown in Figure 7. These two conditions were arguably the best calibrations for these two methods and the results provide evidence that either method works well in calibrating these written expression field test items. In contrast, the M3PL calibrations with mathematics and vocabulary items tended to diverge significantly from those calibrated with written expression. This suggests that even though different ability latent variables for each subject group was specified with the M3PL models, they were still unable to adequately account for the multidimensional nature of the items.

The vocabulary field test item parameter estimates (see Figure 8) were more homogeneous compared to the written expression field test items already discussed. This is particularly true for the pseudo-guessing and discrimination parameters. This result was also shown for the U3PL models discussed above and shown in Figure 4 and may represent items that were easier to calibrate. Similar to the written expression items, the M3PL calibrations using vocabulary operational items produced similar estimates for many of the ten items. Interestingly, there were a few items that diverged significantly between M3PL and U3PL models using vocabulary operational items (see items 1, 2, and 5). In some cases the parameter estimates were consistent for all the estimates except the M3PL model with vocabulary items (see item 5). In this scenario, the more consistent estimates from the U3PL model with vocabulary items may be more desirable to researchers. Lastly, the M3PL models using mathematics and written expression items tended to produce larger estimates for the discrimination and pseudo-guessing parameters, suggesting these may not be the best models for these data.

The results from the M2PL and U2PL analyses (not shown graphically) produced more consistent estimates compared to the 3PL models, particularly for the written expression field test items. The difficulty estimates for the field test items were very consistent across the six models. The discrimination estimates were clustered into two groups, the first represented the mathematics groups and vocabulary MIRT and the second being the vocabulary UIRT and the written expression groups. The first group tended to be estimated as having larger discrimination compared to the second group. This suggests that the best models for this data use written expression as the operational base group.

The vocabulary field test items had more variation in parameter estimates across the six models, but still more consistent compared to the 3PL models. The largest differences were found when estimating the discrimination parameters. The most divergence in parameter estimates occurred between the mathematics and written expression M2PL models and were consistently estimated larger than the rest. Parameter estimates were most consistent when using vocabulary items as the base items and tended to produce the smallest estimates. This again suggests that using vocabulary operational items when calibrating vocabulary field test items produces similar results whether MIRT or UIRT are used.

Discussion

This paper aimed to explore calibration options when a given field test "booklet" is formed from a combination of item types. In this case, the item types consisted of 70% written expression items and 30% vocabulary items. Written expression and vocabulary items theoretically may be very similar, but may represent differences in which there may be an ability unique to each domain. This structure can violate traditional IRT methods that assume that the response strings are a function of a single ability domain (De Ayala, 2013). Through this, the primary research question was to explore how different calibration methods influence parameter estimation of the field test items. The focus was kept at the item level as these are commonly of most interest when developing a new test form.

Through the use of CFA, the structure of the field test items was confirmed to have two unique characteristics (see Table 3). However, the correlation between the two domains was strong, about 0.85 for each booklet. This suggests that the items are contributing unique information, however they are very strongly related. Therefore, the unidimensionality assumption may not be strongly violated and traditional IRT methods assuming unidimensionality may provide adequate estimates. These questions were explored in more detail by comparing different base items used in calibration as well as MIRT methods that allow the researcher to model more than one source of ability estimate.

Results showed that estimates can vary based on which items are used as the base items for calibration. Particularly using mathematics items, representing a severe violation of the unidimensionality assumption, produced parameter estimates that were very different for the field test items compared to other methods with less severe dimensionality assumption violations (e.g. written expression base items). This is not surprising given prior literature surrounding the impact of multidimensionality on parameter estimates (Drasgow & Parsons, 1983; Harrison, 1986; Reckase, 1979). However, much smaller differences were found when the base items represented similar content domains.

MIRT models were also fitted to the data that in theory should allow for the proper estimation of item parameters given multiple ability domains underlying the data (Reckase, 2009). Unfortunately, M3PL models did not perform well unless the base items and the field test item subject matched. Instances in which the item subjects did not overlap, estimates varied widely. Some of this was alleviated by simplifying the model to a M2PL model, an idea found in the UIRT literature (Thissen & Wainer, 1982), however there were still small differences when there was subject mismatch. This may suggest that field test data like those shown here do not satisfy the sample size requirements for MIRT to provide useful estimates. This is an area that future research could inform further.

Implications for Calibrations

The impact of IRT parameter estimates with respect to sample size, number of items, model type, and the unidimensionality assumption is complex. This study did attempt to explore these interactions with real data when there is small evidence of multidimensionality with field test items. In this framework, similar estimates were obtained as long as the base operational items used in the calibration are related to either of the two overlapping domains, for example using either written expression or vocabulary base items to calibrate written

expression or vocabulary field test items. This was true for both U3PL and U2PL models, however the estimates were much more precise for U2PL models compared to U3PL models an idea explored by others (Author, Under Review; Thissen & Wainer, 1982). These differences can yield very different ICCs for items, which may impact the overall test characteristic curve if the estimation is consistently over- or under-estimated over many items, which could influence scores on the assessments.

Implications for Field Test Decision Making and Design

When collecting data from a field test, sample sizes can at times be small. This has direct implication for the quality of results obtained from calibrations and is not a new idea in psychometric research (Author, Under Review; Thissen & Wainer, 1982). This result was further found in this current study with regard to the amount of variation in the calibration methods and the size of the standard errors (shown by the confidence interval bands), especially for the pseudo-guessing parameter. One way to improve estimation precision is to simplify the model from a 3PL to a 2PL model. This result held for both MIRT and UIRT models and can increase the confidence and usefulness of the calibration results.

This change has at least two implications for making informed decisions regarding which field test items are best used on an operational form. First, simplifying the model can directly influence the quality and consistency of parameter estimates across models. This can be important as the U2PL models were more robust to sample size constraints and violations to the unidimensionality assumption. This was not true of the 3PL models under either the MIRT or UIRT framework. Secondly, simpler models will allow for less concern regarding sample sizes for test developers trying out field test items. More items could be tried out or tried out a second time with less concern about variability in the estimated parameters.

References

Ackerman, T. A. (1994). Using multidimensional item response theory to understand what items and tests are measuring. *Applied Measurement in Education*, *7*(4), 255–278.

Author. (Under Review). Validity of the three parameter logistic item response theory model for field test data.

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

Chalmers, R. P. (2012). Mirt: a multidimensional item response theory package for the r environment. *Journal of Statistical Software*, *48*(6), 1–29.

Chen, W.-H. & Thissen, D. (1997). Local dependence indexes for item pairs using item response theory. *Journal of Educational and Behavioral Statistics*, *22*(3), 265–289.

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

DeMars, C. E. (2013). A tutorial on interpreting bifactor model scores. *International Journal of Testing*, *13*(4), 354–378.

Downing, S. M. & Haladyna, T. M. (1997). Test item development: validity evidence from quality assurance procedures. *Applied Measurement in Education*, *10*(1), 61–82.

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

Drasgow, F. & Parsons, C. K. (1983). Application of unidimensional item response theory models to multidimensional data. *Applied Psychological Measurement*, *7*(2), 189–199.

Finch, H. & Habing, B. (2007). Performance of dimtest-and noharm-based statistics for testing unidimensionality. *Applied Psychological Measurement*, *31*(4), 292–307.

Green, S. B. & Yang, Y. (2015). Evaluation of dimensionality in the assessment of internal consistency reliability: coefficient alpha and omega coefficients. *Educational Measurement: Issues and Practice*, *34*(4), 14–20.

Hambleton, R. K., Jones, R. W., & Rogers, H. J. (1993). Influence of item parameter estimation errors in test development. *Journal of Educational Measurement*, *30*(2), 143–155.

Harrison, D. A. (1986). Robustness of irt parameter estimation to violations of the unidimensionality assumption. *Journal of Educational and Behavioral Statistics*, *11*(2), 91–115.

Henning, G., Hudson, T., & Turner, J. (1985). Item response theory and the assumption of unidimensionality for language tests. *Language testing*, *2*(2), 141–154.

Holzinger, K. J. & Swineford, F. (1937). The bi-factor method. *Psychometrika*, *2*(1), 41–54.

Jones, P., Smith, R. W., & Talley, D. (2006). Developing test forms for small-scale achievement testing systems. *Handbook of test development*, 487–525.

Kim, S.-H. & Cohen, A. S. (1991). A comparison of two area measures for detecting differential item functioning. *Applied psychological measurement*, *15*(3), 269–278.

Kline, R. B. (2011). *Principles and practice of structural equation modeling*. Guilford Press.

Kolen, M. J. & Brennan, R. L. (2004). *Test equating, scaling, and linking*. Springer.

Livingston, S. A., Downing, S., & Haladyna, T. (2006). Item analysis. *Handbook of test development*, 421–441.

Lord, F. M. (1968). An analysis of the verbal scholastic aptitude test using birnbaum's three-parameter logistic model. *Educational and Psychological Measurement*.

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

McDonald, R. P. (2013). *Test theory: a unified treatment*. Psychology Press.

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

Oshima, T. & Miller, M. D. (1992). Multidimensionality and item bias in item response theory. *Applied Psychological Measurement*, *16*(3), 237–248.

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/

Raju, N. S. (1988). The area between two item characteristic curves. *Psychometrika*, *53*(4), 495–502.

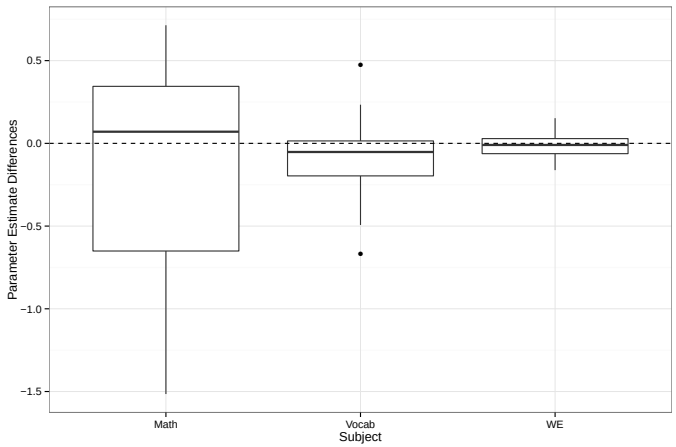
Reckase, M. (1979). Unifactor latent trait models applied to multifactor tests: results and implications. *Journal of Educational and Behavioral Statistics*, *4*(3), 207–230.

Reckase, M. (2009). *Multidimensional item response theory*. Springer.

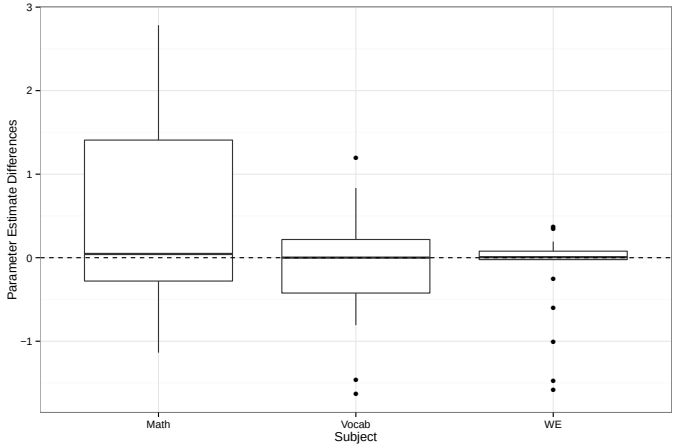
Stout, W. (1987). A nonparametric approach for assessing latent trait unidimensionality. *Psychometrika*, *52*(4), 589–617.

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

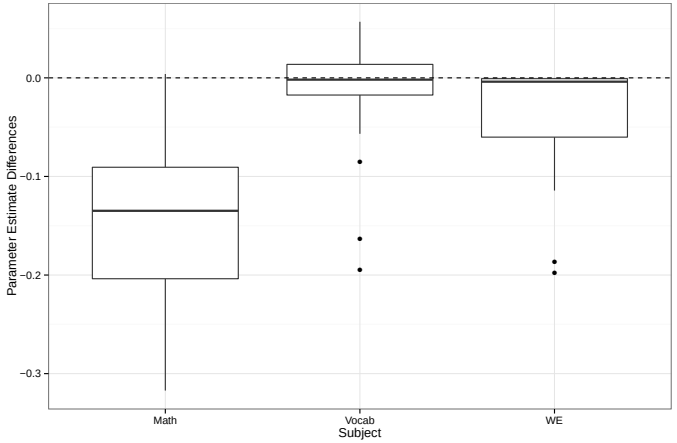
Wickham, H. (2009). *Ggplot2: elegant graphics for data analysis*. Springer New York. Retrieved from http://had.co.nz/ggplot2/book



(a) Discrimination

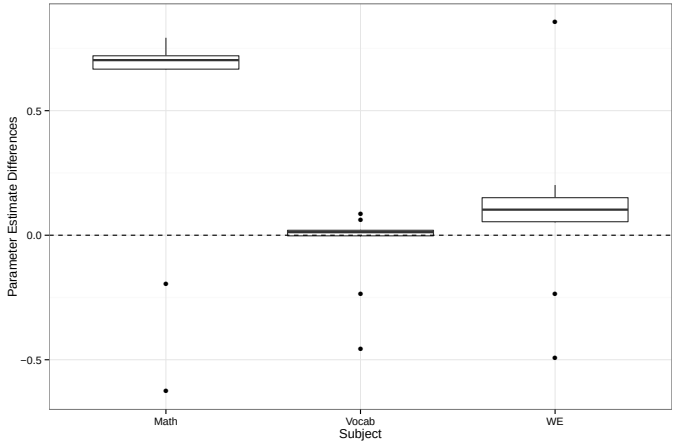


(b) Difficulty

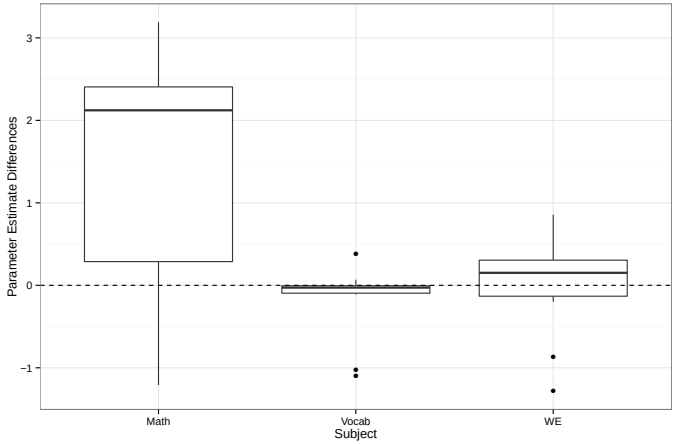


(c) Pseudo-guessing

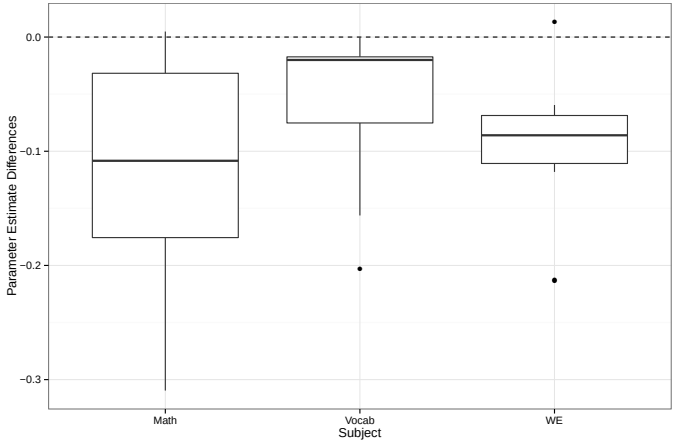
Figure 1: Box plots of differences in parameter estimates using a 3PL model with written expression field test items



(a) Discrimination

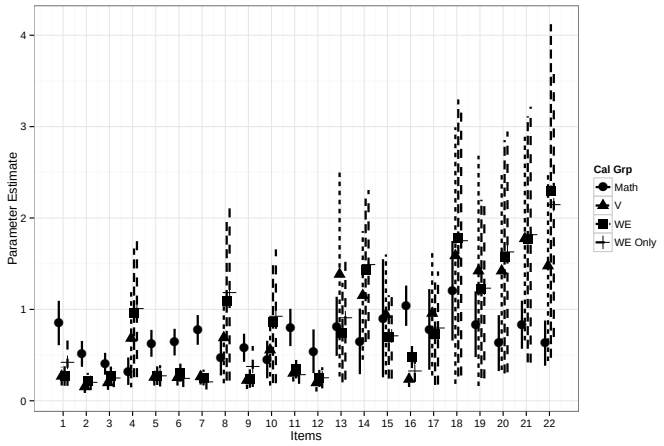


(b) Difficulty

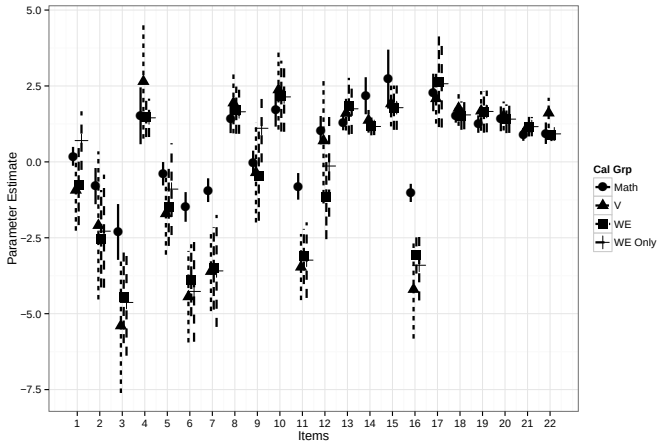


(c) Pseudo-guessing

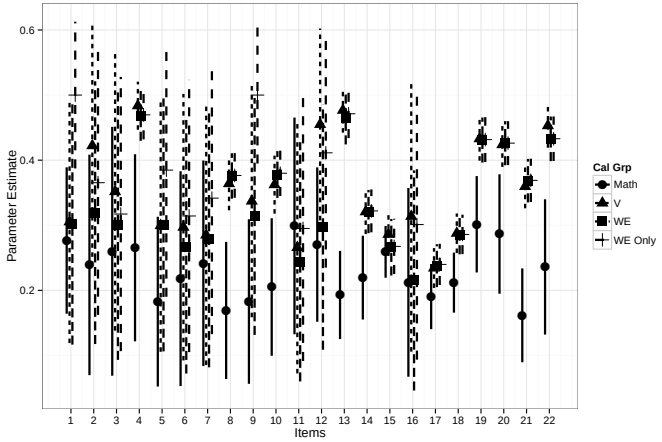
Figure 2: Box plots of differences in parameter estimates using a 3PL model with vocabulary field test items



(a) Discrimination

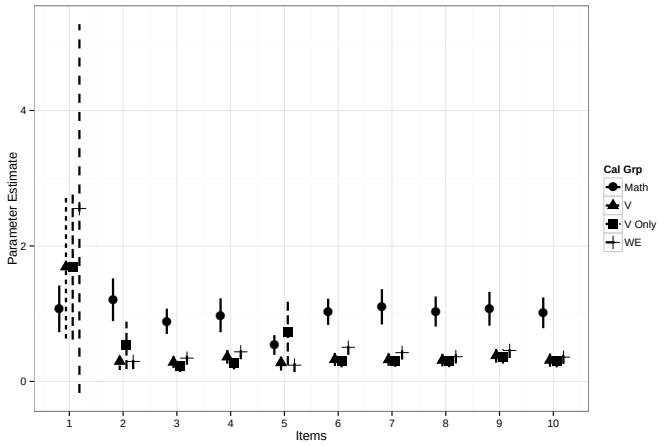


(b) Difficulty

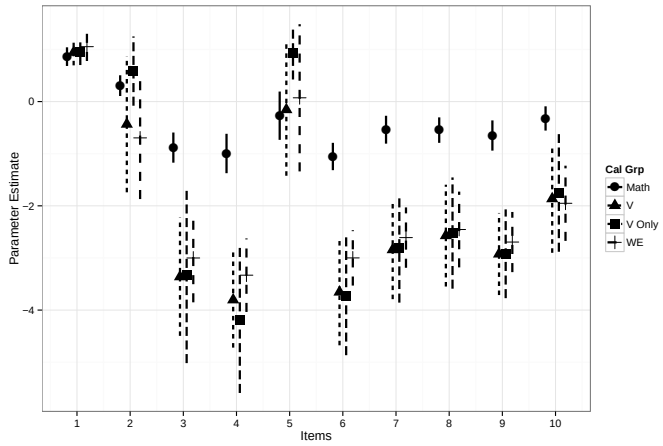


(c) Pseudo-guessing

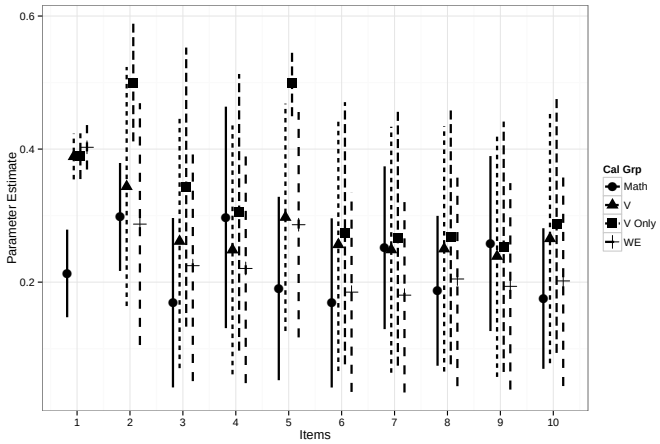
Figure 3: 95% confidence intervals of parameter estimates using different base items for calibration with the 3PL model with written expression field test items



(a) Discrimination

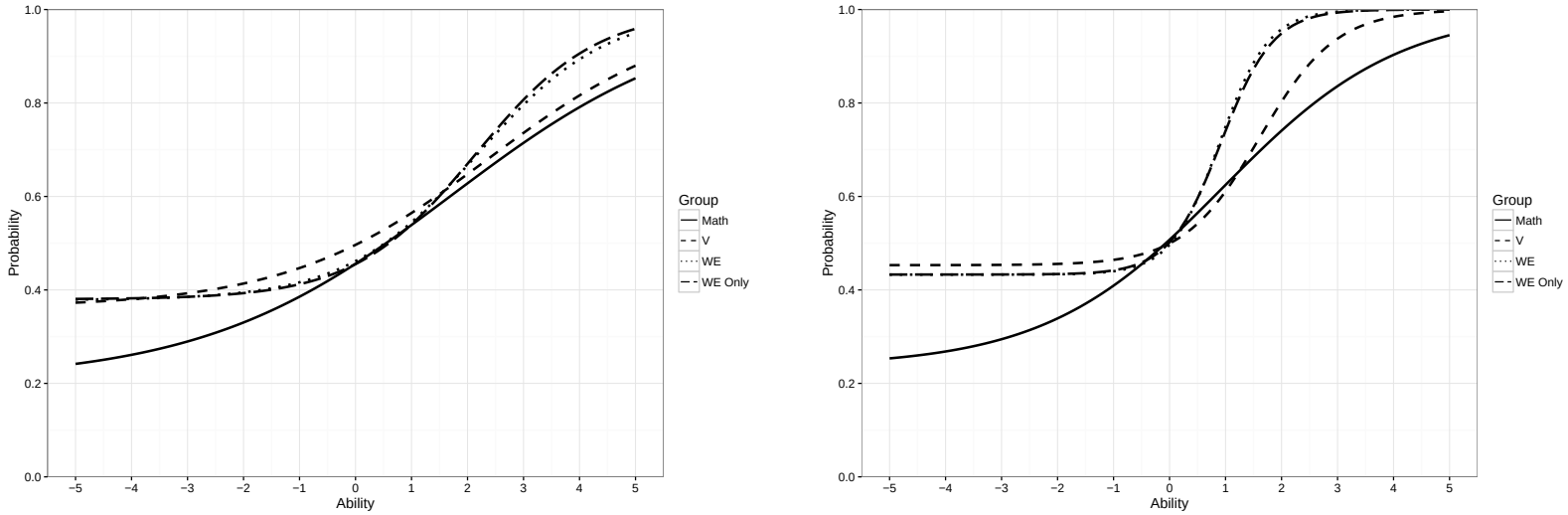


(b) Difficulty



(c) Pseudo-guessing

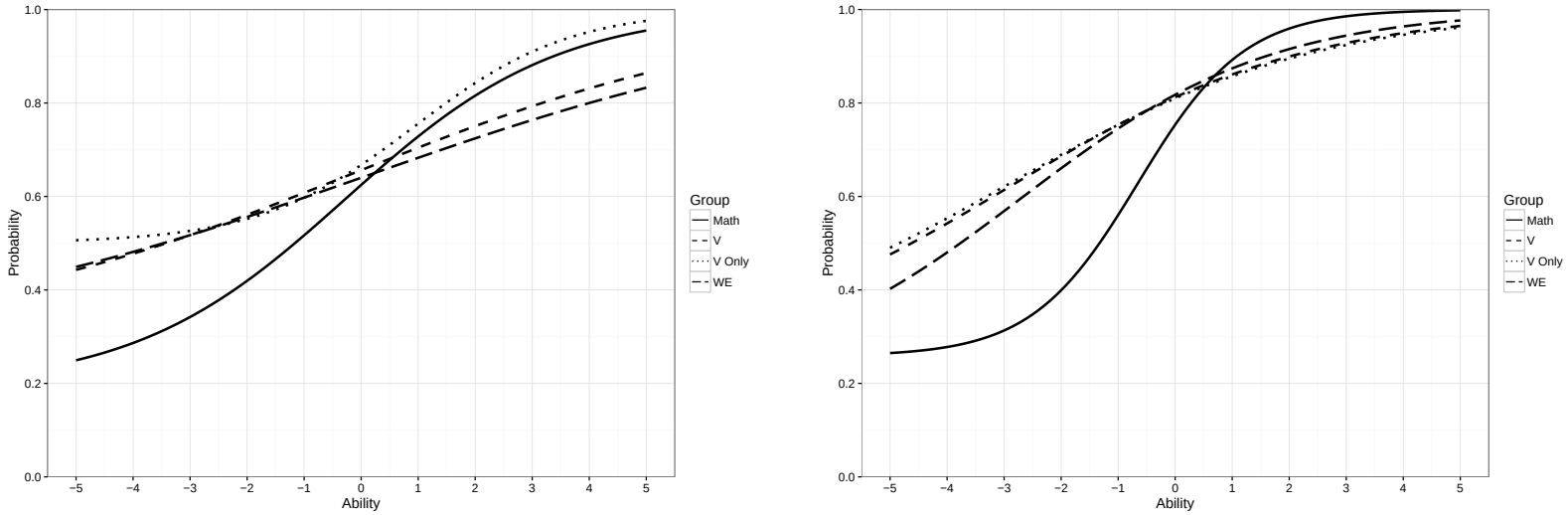
Figure 4: 95% confidence intervals of parameter estimates using different base items for calibration with the 3PL model with written expression field test items



(a) Item 10

(b) Item 22

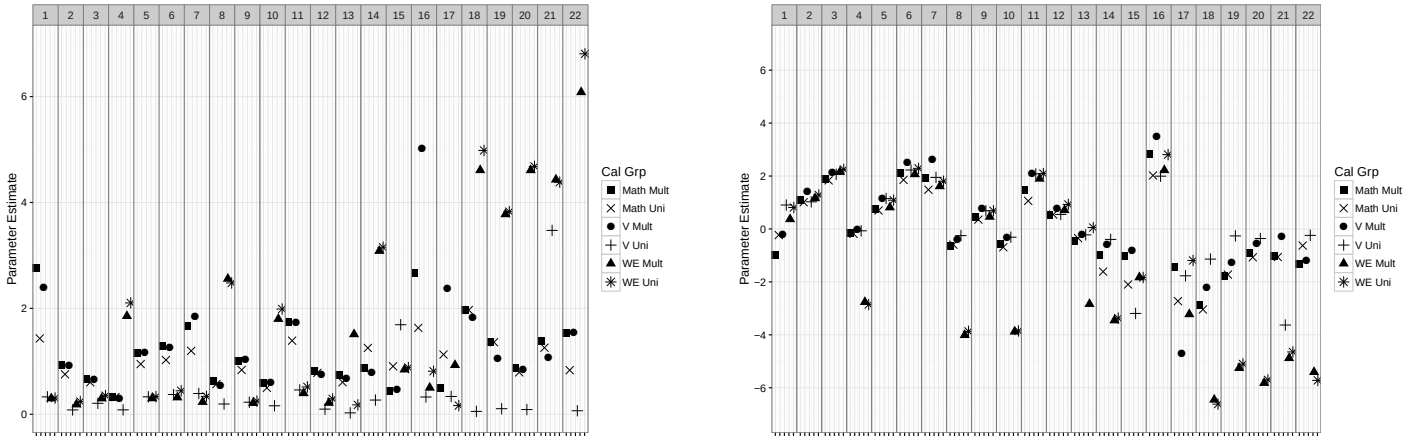
Figure 5: Item characteristic curves for two 3PL written expression field test items



(a) Item 5

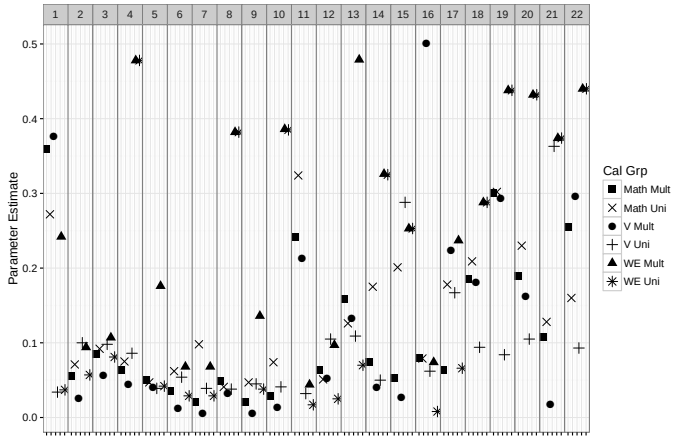
(b) Item 9

Figure 6: Item characteristic curves for two 3PL vocabulary field test items



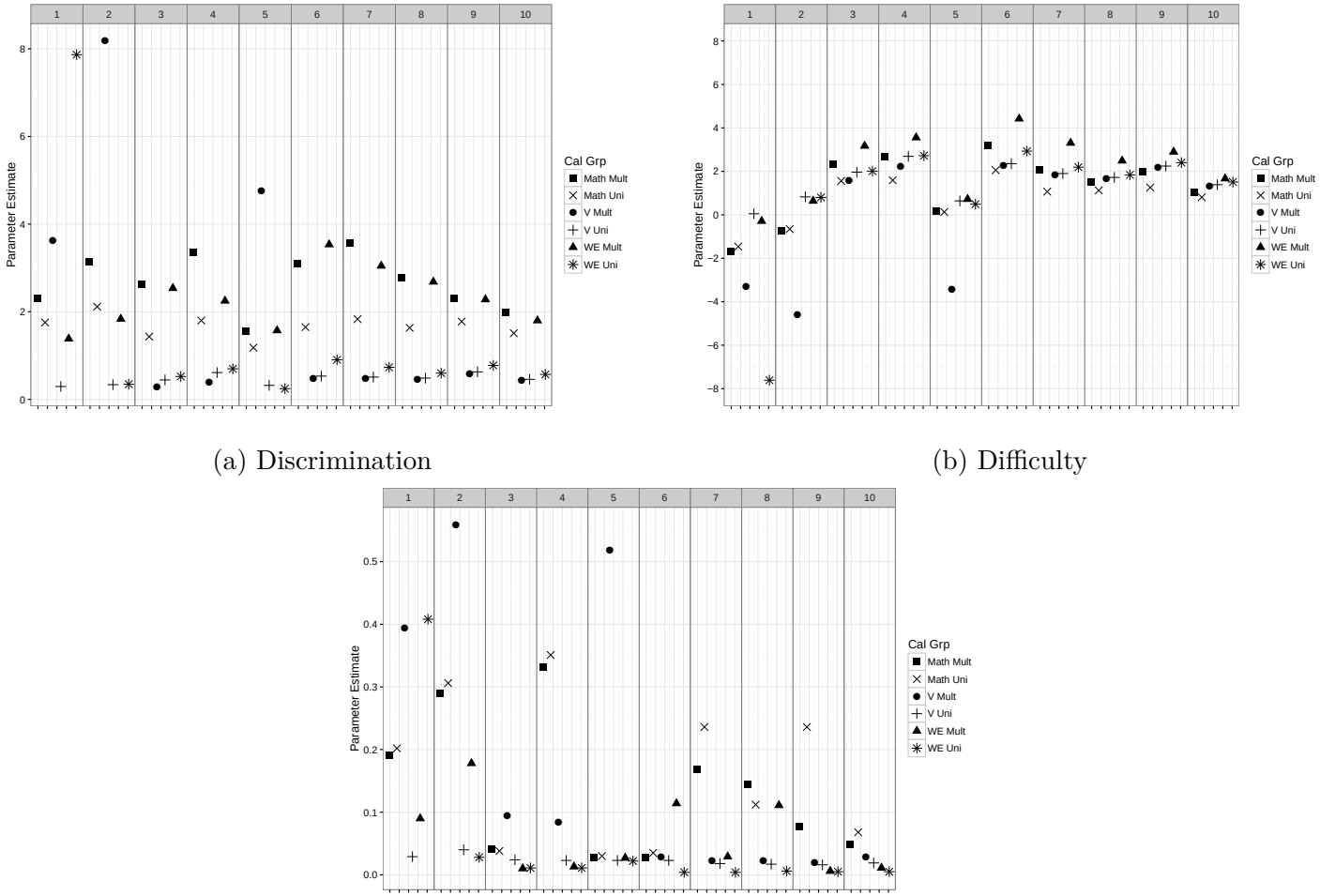
(a) Discrimination

(b) Difficulty



(c) Pseudo-guessing

Figure 7: Parameter estimates for written expression field test items comparing multidimensional and unidimensional IRT with three different sets of base items.



(a) Discrimination

(b) Difficulty

(c) Pseudo-guessing

Figure 8: Parameter estimates for vocabulary field test items comparing multidimensional and unidimensional IRT with three different sets of base items.