APPENDIX 6-C. 2015 NCSC Item Response Theory Model Selection

This document provides a high-level summary of the preliminary results of the 2015 NCSC operational data analyses. Measured Progress Psychometrics' recommendations on selecting the item response theory (IRT) model will follow.

As described in the *2015 NCSC Operational Analysis Plan*, Measured Progress fitted both Rasch and 2-parameter logistic (2PL) models to the operational NCSC data (i.e., 14 grade/content combinations). Evaluation and comparison of the model-data fit will facilitate the determination of the appropriate IRT model. The 2PL model for dichotomous items can be defined as:

$$P_i(X_i = 1|\theta_j) = \frac{\exp[Da_i(\theta_j - b_i)]}{1 + \exp[Da_i(\theta_j - b_i)]}$$

where
X_i indexes the scored response on an item,
i indexes the items,
j indexes students,
α represents item discrimination,
b represents item difficulty,
D is a normalizing constant equal to 1.701.

For the Rasch model, the above equation reduces to the following:

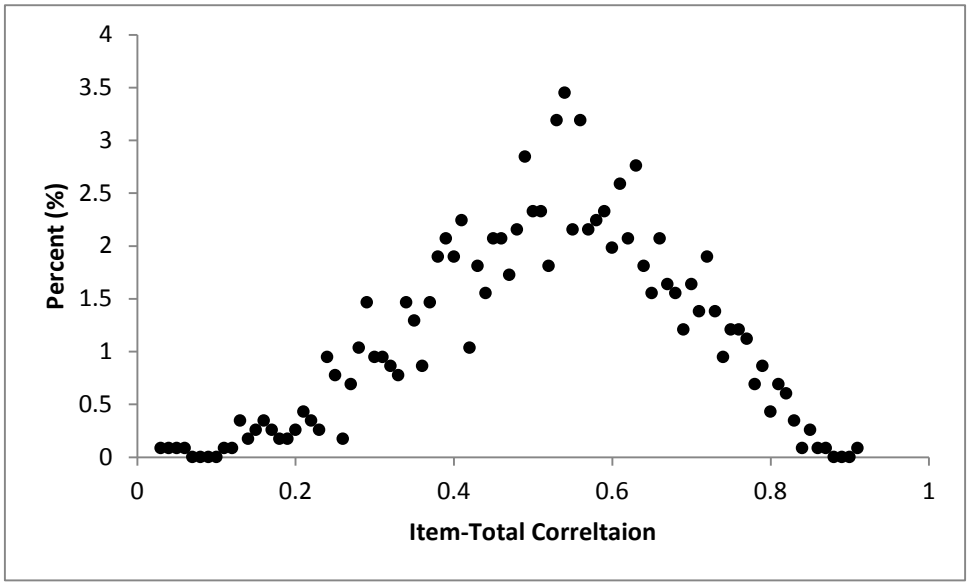
$$P_i(X_i = 1|\theta_j) = \frac{\exp(\theta_j - b_i)}{1+\exp(\theta_j - b_i)}.$$

The main difference between Rasch and 2PL models is whether the theoretical model assumes equal item discrimination across all items or not. That is, the Rasch model only estimates the item difficulty parameters (i.e., *b*) while assuming a uniform discrimination parameter (i.e., *a*) of 1 for every item on the test, whereas 2PL estimates both the item difficulty and item discrimination parameters for all items. IRT parameters *a* and *b* are closely related to the corresponding classical statistics, *p*-value and item-total correlation, quantifying the item difficulty and discrimination, respectively.

Preliminary Classical and IRT Statistics

As a starting point, Measure Progress Psychometrics examined the classical statistics, specifically focusing on the item-total correlation of all items. If the values are approximately uniform across different items, the Rasch model would result in a better a model-data fit, however, if the values show variability, the 2PL model would better fit the data. The distribution of the item-total correlation values is plotted below in Figure 1 (Note: For all figures provided in this document, 1,160 items from all 14 grade/content area combinations were plotted.).

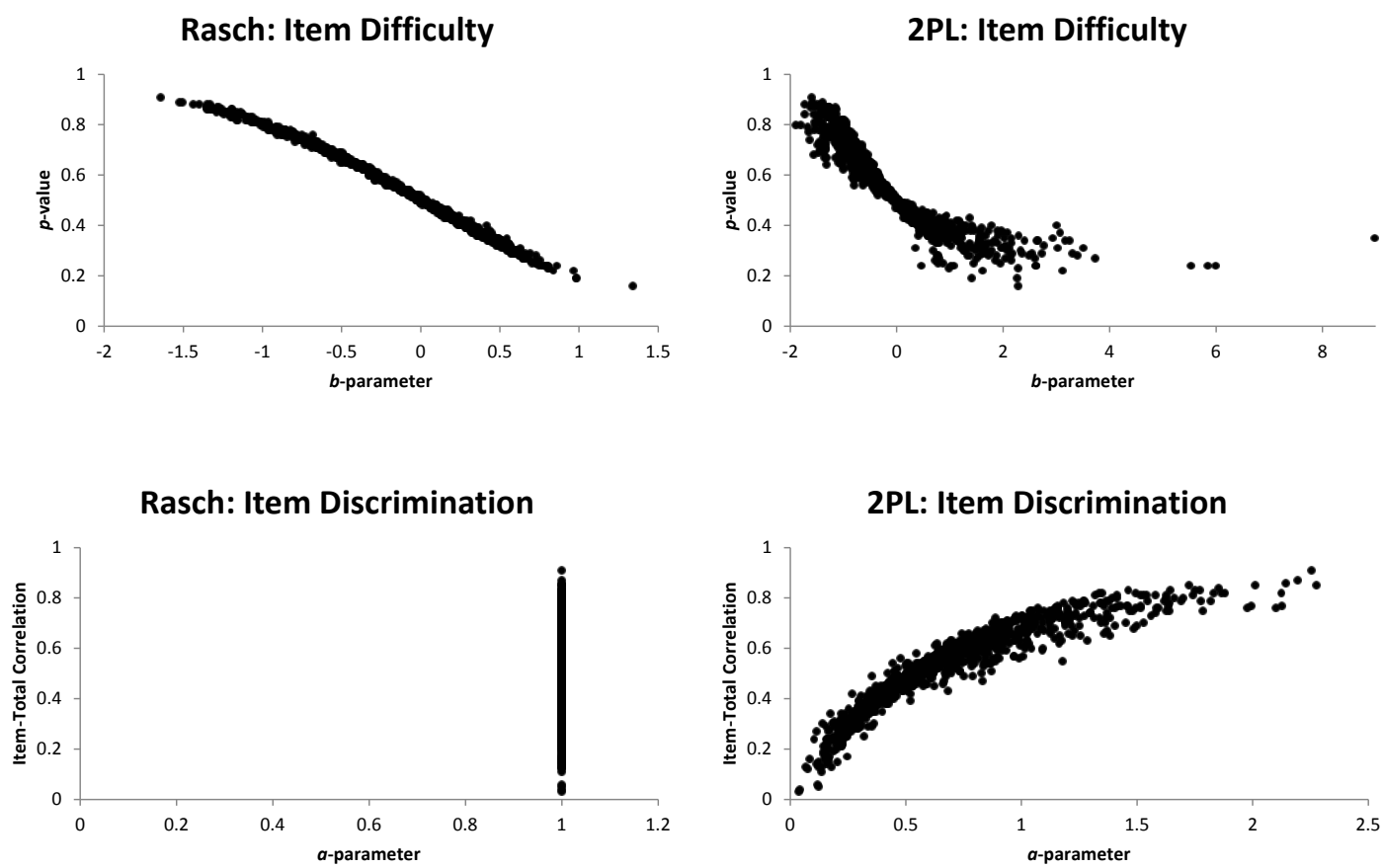
Figure 1. Distribution of Item-Total Correlation



As can be observed in Figure 1, the 2015 NCSC operational data indicate variability in item-total correlation (min=.03, max=.91, mean=.52, and SD=.16), meaning that some items perform better in discriminating the more-abled examinees from the less-abled examinees.

Then the IRT parameters estimated under each model were examined in relation to the corresponding classical statistics (see Figure 2).

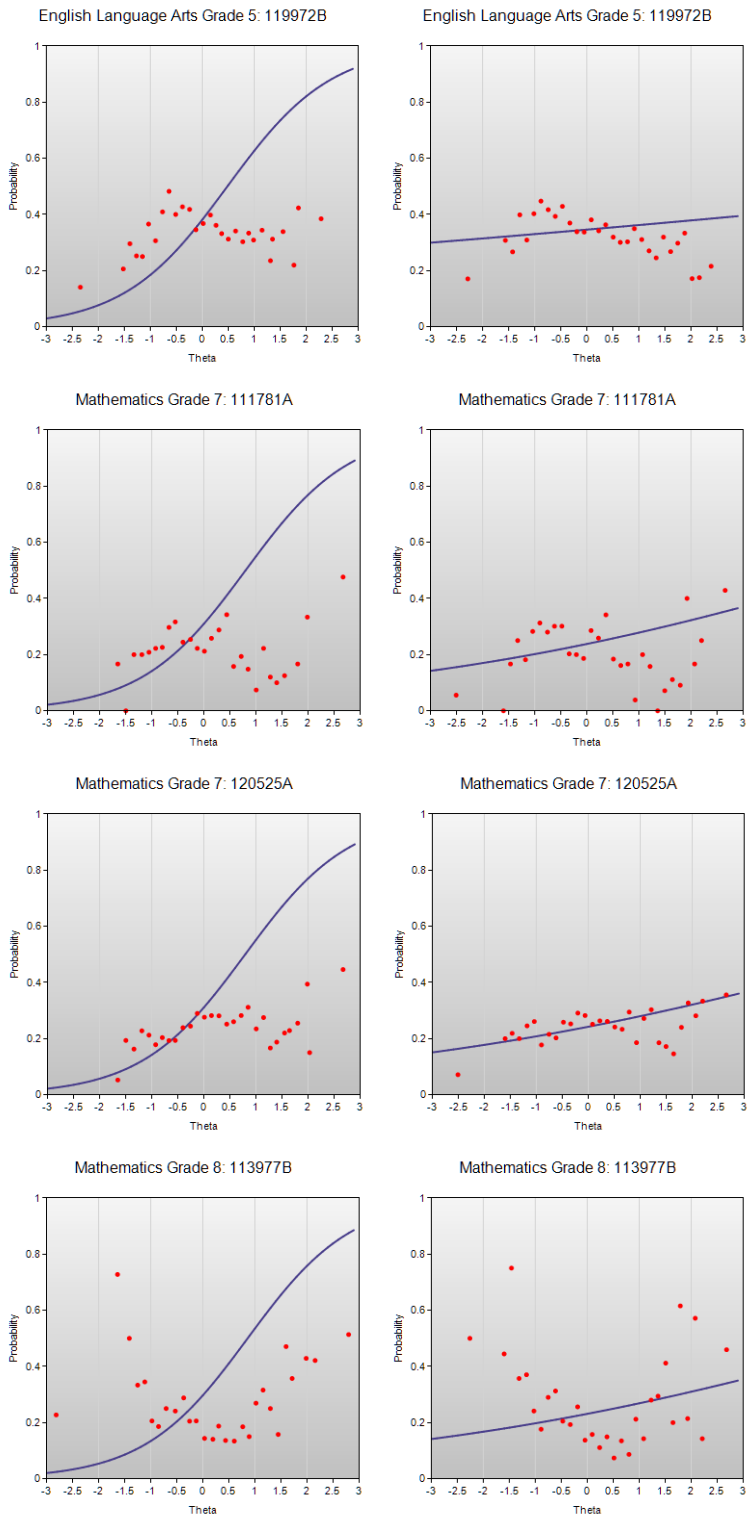
Figure 2. Item Difficulty and Item Discrimination for Classical and IRT Statistics: Rasch v. 2PL



The two plots in the first row of Figure 2 clearly display a *negative* relationship between the classical *p*-values and the IRT *b*-parameters. (These relationships are also reflected in the correlation coefficients: -.99 for Rasch and -.87 for 2PL.) The increase in *p*-values indicates easier items, which translates to lower values in terms of *b*-parameters. The second row of Figure 2 demonstrates the relationship between the classical item-total correlation and the *a*-parameters calibrated under each IRT model. For Rasch, the *a*-parameter is fixed to equal 1 regardless of the classical item-total correlation. However, when the 2PL model is fitted to the data, separate item discrimination parameters are estimated for each item. It can be easily seen from the plot that the relationship between the classical and IRT statistics is *positive*, indicating that the 2PL model takes item-total correlation into account when trying to find a best-fitting curve to the empirical data. (The correlation between item-total correlation and *a* is .91.)

Note that some *b* estimates calibrated under 2PL are extreme (i.e. larger than 4) while none of the *b* parameter estimates showed extreme values when calibrated under Rasch in Figure 2. Those items particularly produced low item-total correlation values, and they were scrutinized in terms of item-level fit plots (see Figure 3). Fit plots for four items are shown in Figure 3 where each row represent each item. The plots on the left resulted from a Rasch run and the ones on the right resulted from a 2PL run. When the Rasch model is fit, the standard error of the *b* parameter is determined by sample size only, which is not informative of the model-data fit. However, in the case of 2PL, the standard error of *b* actually gives us insight into the model-data fit because an item with low discrimination will cause the standard error of *b* to be larger, and *b* estimates to be extreme. Since the 2PL model cannot accurately measure the shape of the item response function (IRF), we see “appropriately” high standard errors of the *b*-parameter estimates.

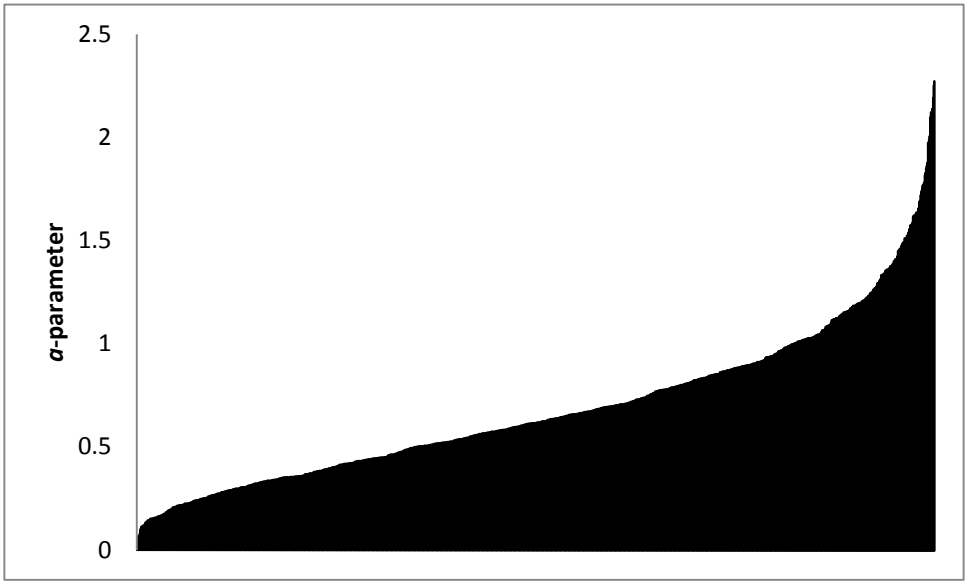
Figure 3. Item-Level Fit Plots for Items with Extreme *b*-parameters



To closely investigate the variability of the IRT item discrimination calibrated under the 2PL model, the *a*-parameter estimates for all items were plotted in Figure 4. If the Rasch model were to best fit the empirical data, we would expect a) the *a*-parameters to demonstrate an approximate uniform distribution around the y-axis value of 1; and b) the mean and the standard deviation of the distribution to be quite close to 1 and 0, respectively. However, as can be observed in Figure

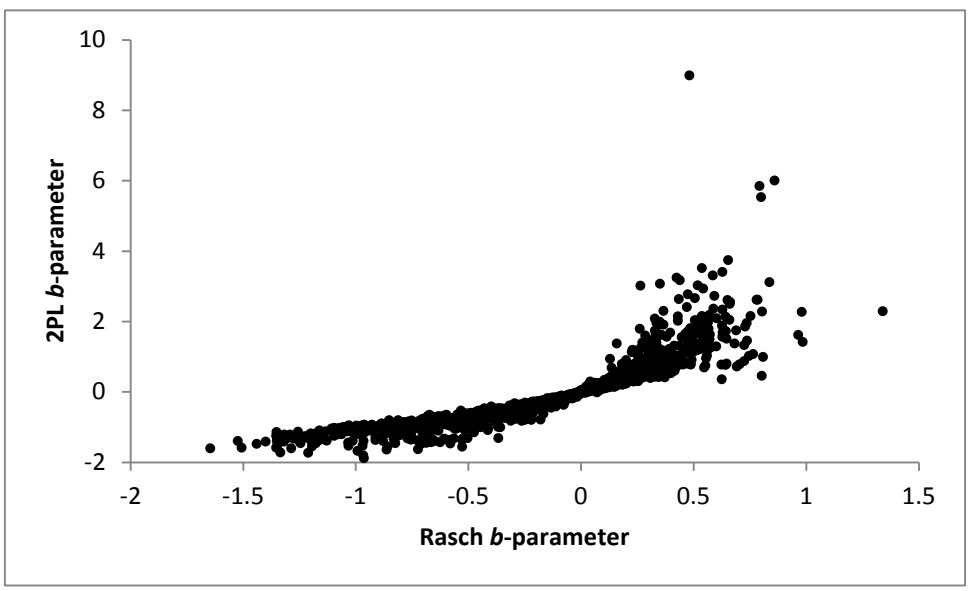
4, the item discrimination estimates ranges from a value close to 0 to a value larger than 2 (with a mean of .69 and a standard deviation of .38).

Figure 4. Distribution of *a*-parameter



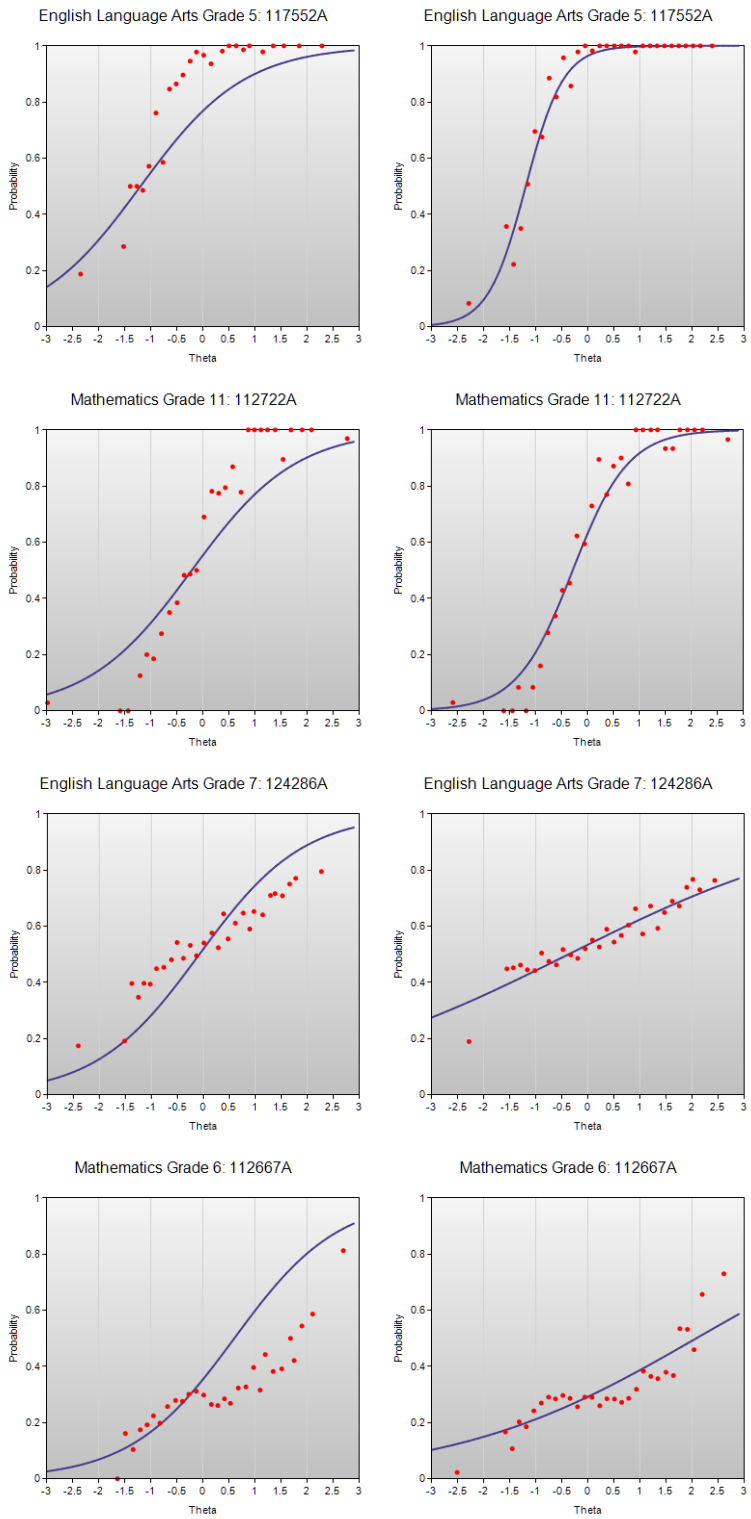
Then the relationship between the *b*-parameters calibrated under Rasch and 2PL were investigated. In Figure 5, we can see that the difficulty parameters have a positive relationship (The correlation between the two *b*s was .86.) indicating that the ordering of the item difficulty estimates does not hugely rely on the choice of IRT models.

Figure 5. Relationship between Rasch and 2PL *b*-parameters



Next, individual item-level fit plots were examined. In Figure 6, example item-level fit plots for four items are provided: each from grades 5 and 7 ELA, and grade 6 and 11 mathematics, respectively. The first two rows give examples of high item-total correlation items, while the last two rows give examples of low item-total correlation item. In all figures, the plots on the left resulted from a Rasch run and the ones on the right resulted from a 2PL run.

Figure 6. Item-Level Fit Plots for Rasch and 2PL for Items with High and Low Item-Total Correlations



From examining Figure 6, we can see that the fit between the model (dark blue lines) and the empirical data (red dots) improved by using 2PL instead of Rasch. The first two example items showed relatively high item-total correlation coefficients (.81 for item 1175523A and .67 for item 112722A), and the last two example items showed relatively low item-

total correlation coefficients (.31 for item 124286A and .29 for item 112667A). While the 2PL model utilized the fact that the first two items discriminated the low- and high-performing examinees quite well and the last two items did not, the Rasch model ignored the high or low item-total correlations and forced the item discrimination parameter *a* to equal 1 for all of these items resulting in unsatisfactory model-data fits.

Recommendations

By carefully scrutinizing classical statistics and preliminary IRT results, Measured Progress Psychometrics recommends the use of the 2PL model to process the operational data for NCSC. The reasons for recommending 2PL are:

- The distribution of item-total correlation indicates variability in item-total correlation.
- Test developers usually focus on both *p*-values and item-total correlation when they construct test forms, and 2PL utilizes both item difficulty and discrimination parameters that correspond to the classical statistics used for test construction.
- NCSC has adequate sample size to fit the 2PL model to the data (~4,000 examinees per grade), and the IRT software did not present any convergence issue when the 2PL model was fitted.
- Model-data fit improves with the use of 2PL especially for items with high or low item-total correlation.
- NCSC will likely lose more items due to bad model-data fit if the Rasch model is used to process the data.
- With the use of 2PL, the predicted reliability (i.e., test information function or standard error of scale scores) from the IRT analyses will better match the classical test theory based reliability coefficient.