APPENDIX 6-D. 2015 NCSC Scoring Decisions

This document provides a high-level summary of the preliminary results of the 2015 NCSC operational data analyses. Measured Progress Psychometrics' recommendations on scoring special item types (i.e., foundational items, multi-part selected response items, and Tier 1 writing prompt items) will follow.

Following the decision made on the item response theory (IRT) modeling decision, Measured Progress Psychometrics fitted the 2-parameter logistic (2PL) model to the operational NCSC data (i.e., 14 grade/content combinations). For special item types, two different scoring approaches can be taken and affect the IRT analyses:

- "Combined" approach: To assume local dependence and combine the item scores and assign a score point following the scoring rules that were used when Pilot 2 was administered; or
- "Separate" approach: To assume local independence and treat each item as an independent selected response (SR) item.

Measured Progress carefully examined the special item types and analyzed the student responses using both approaches.

Foundational Items

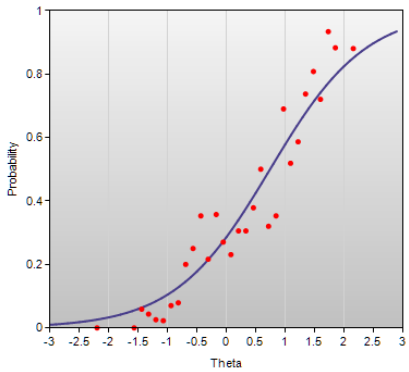
Foundational items appear in the grades 3 and 4 ELA assessments. They consist of either three or five dichotomously (i.e., 0/1) scored SR items. For Pilot 2, the rules for assigning final scores for the foundational items were (hereinafter referred to as the "combined" approach):

- For Tier 1 foundational items, a student had to score 3 out of 3 to obtain a score of 1; otherwise, a score of 0 was assigned.
- For Tiers 2 through 4 foundational items, a student had to score 4 or 5 out of 5 to obtain a score of 1; otherwise, a score of 0 was assigned.

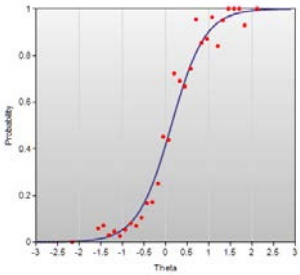
First, Measured Progress fitted the 2PL model to the empirical data based on the "combined" approach. Then, 2PL was also fitted to the separately 0/1 scored responses ("separate" approach). Below are some example item-level fit plots for the two different scoring approaches (see Figure 1). The plots on the left represent the "combined" approach whereas the plots on the right represent the "separate" approach.

Figure 1. Example Item-Level Fit Plots for Foundational Items

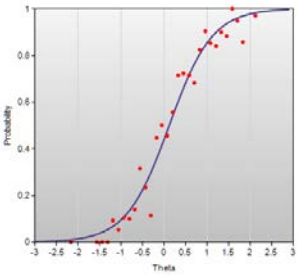
English Language Arts Grade 3: 300011V



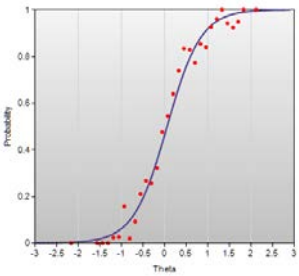
English Language Arts Grade 3: 121592B



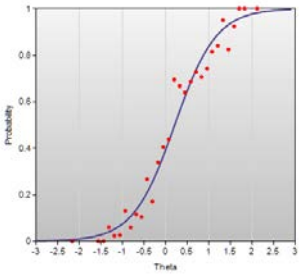
English Language Arts Grade 3: 121609B



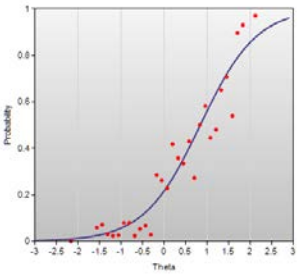
English Language Arts Grade 3: 121612A



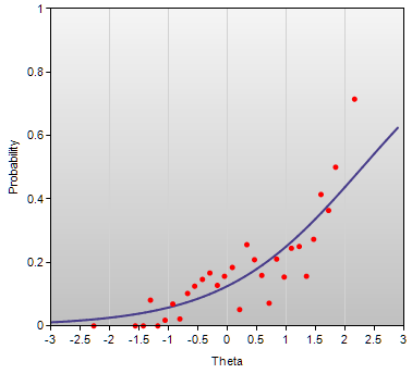
English Language Arts Grade 3: 121594A



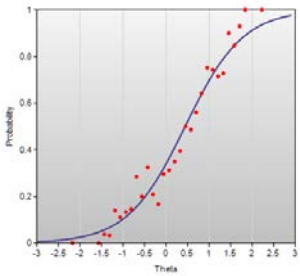
English Language Arts Grade 3: 121611A



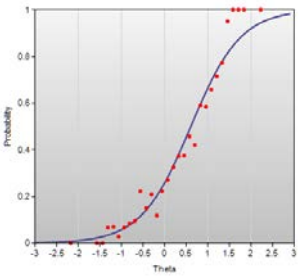
English Language Arts Grade 4: 400014V



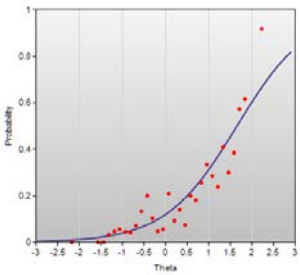
English Language Arts Grade 4: 121618A



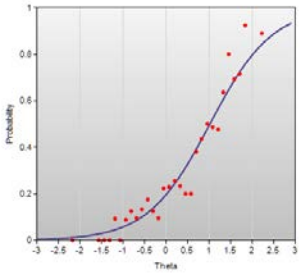
English Language Arts Grade 4: 121620A



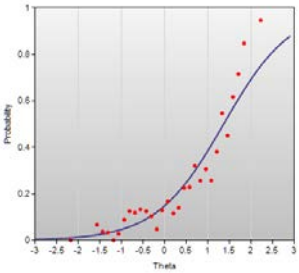
English Language Arts Grade 4: 121622A



English Language Arts Grade 4: 121619A



English Language Arts Grade 4: 121621A



As can be seen in the right column in Figure 1, each item displayed different shapes of item response functions (IRFs). These different shapes, supported by the psychometric estimates and the empirical data, indicate that the separate items vary in both difficulty and discrimination. Also, for both items presented above, out of the five items, some items show better model-data fit than the others. Combining the scores (the "combined" approach) did not meaningfully improve the model-data fit. In addition, there is evidence to support that an unnecessary increase in difficulty is introduced to the foundational items by using the "combined" approach, meaning that it is much more difficult for an examinee to get a full credit on these items, without gaining significant improvement in terms of psychometric characteristics.

Foundational items are surely measuring the same underlying construct, and a valuable one, too. These items are independently presented to examinees (i.e., not tied to a specific reading passage), and scored independently as 0/1's, and the score obtained on a previous item does not affect the score the student obtains on the next item. Hence, when fitting the IRT model to the data, separately modeling individual items is quite sensible. By fitting 2PL to each item separately, 1) we will be able to more accurately estimate the mathematical shape of the IRF that best-fits the student responses; 2) we can better evaluate each individual item to identify items that fit the IRT model better or flag the items that do not show satisfactory model-data fit; 3) when we lose an item due to a flaw or if we need to replace an item for some reason, the decision can be made on an item by item basis instead of for the entire three- or five item-set; 4) it will be easier to edit and revise challenging item types and ultimately improve the item development process for this population of students; and 5) it will be easier to select “linking” items in the future and will make the implementation of the staged-adaptive model more flexible (if desired).

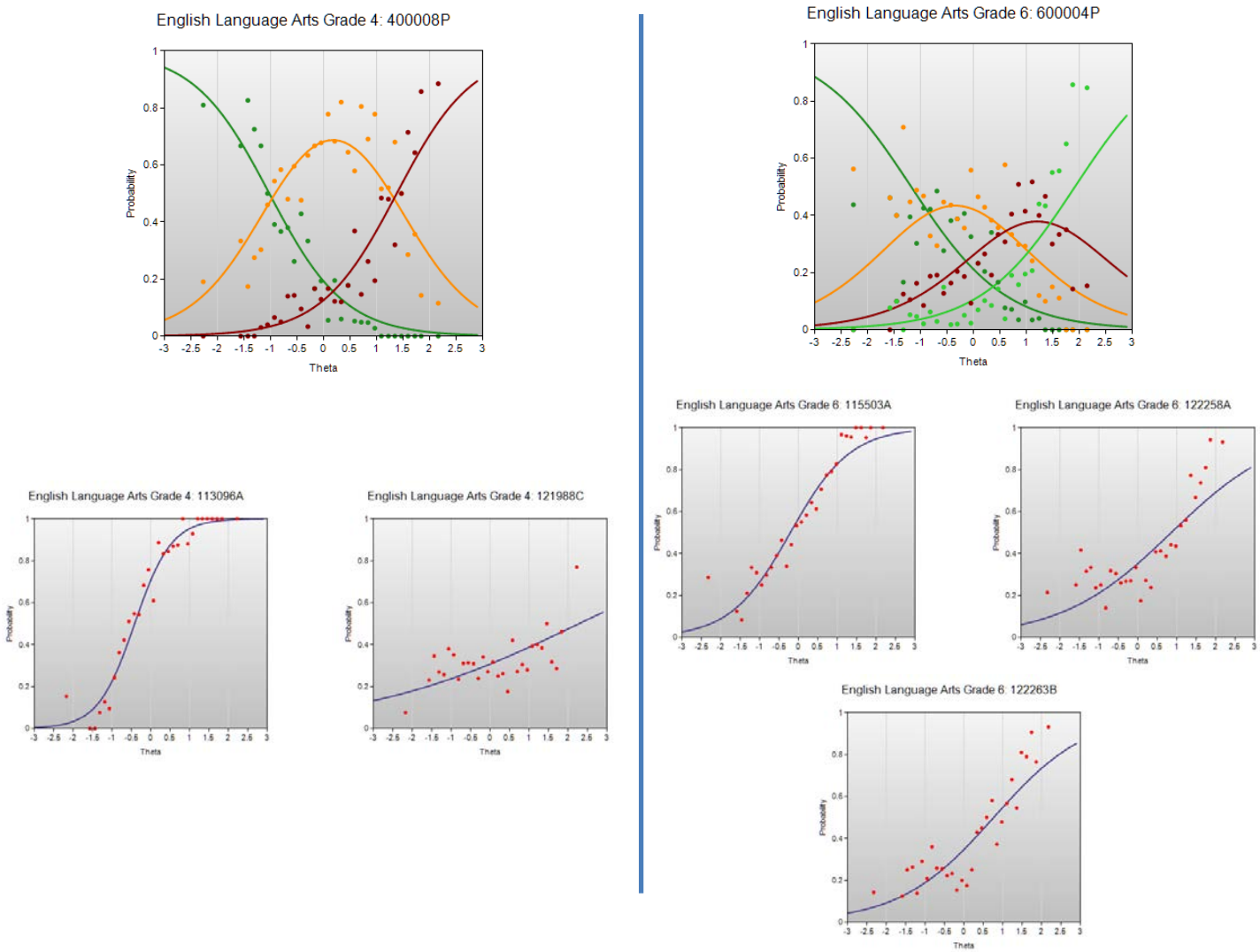
Multi-Part Selected Response Items

Multi-part selected response (SR) items were included in all ELA test forms regardless of the grade. They consisted either of two or three independently (and dichotomously) scored SR items. The correct answer to the previous item is always stated so that the response to the previous item does not have any influence on the probability of an examinee correctly answering the subsequent item(s). For Pilot 2 (“combined” approach), the final scores given to the students reflected the number of SRs they answered correctly out of those two or three items (i.e., possible scores of 0, 1, or 2 for the 2-part SR items and 0, 1, 2, or 3 for the 3-part SR items).

Measured Progress fitted the graded response model (GRM), an extended version of the 2PL model used for polytomous responses, to the “combined” data and the 2PL model to the “separate” data. Some example fit plots are presented in Figure 2.

In Figure 2, each of the two upper plots displays multiple curves representing the “combined” approach for two example cases. Also shown in the Figure 2 below each multi-curve plot are two sets of single-curve plots representing the “separate” approach for each case (two plots for one case and three for the other). These two cases are illustrative of poor model-data fit for the “combined” approach, indicating that when the items are scored using the “combined” approach, adequate psychometric characteristics may not result. Such misfit could be due to problematic behavior by either all the items in a set or only a subset of the items. As can be easily seen in the lower portions of Figure 2, by calibrating each SR item separately we are able to clearly see which items out of two (or three) are exhibiting the problematic behavior. Each item display its own unique characteristics in terms of model-data fit, item discrimination, and item difficulty, etc.; and this information is critical for evaluating the psychometric quality of the assessment.

Figure 2. Example Item-Level Fit Plots for 2-Part and 3-Part Selected Response Items



As discussed earlier, better model-data fit was observed for some items than others within the multi-part SRs, and the fit was not improved when the scores were combined into a single polytomously scored item. Each item displayed different shapes of item response functions (IRFs) indicating that the empirical data show varying item discrimination and item difficulty for the items before they were combined. By using the "combined" approach important psychometric information is lost, and the IRT results become less accurate than using the "separate" approach. The varying item difficulty and discrimination values are masked with the polytomous IRT model and psychometric assumptions necessary for a valid application of a polytomous IRT model are violated.

These items are independently scored as 0/1's, and the item responses are not dependent psychometrically. As with foundational items, fitting the 2PL model to each individual item seems reasonable. The "separate" approach will let us 1) obtain more accurate IRT estimates; 2) evaluate individual items without losing any psychometric information; 3) make decisions based on each item instead of the two- or three-item sets; and 4) better isolate the items that are not functioning well and offer the potential for researching the impact of the content revisions and ultimately may help us better measure the knowledge skills and abilities of this student population.

For example, for the 2015 operational NCSC, one item from a 2-part SR item was flawed due to an error in the audio file. This item will definitely have to be removed from the core item list. If we were to use the "combined" approach, the decision will apply to both items that combine to form that 2-part SR item. However, if a "separate" approach is used, we can make a decision based on each item, and decide to remove only one of the 2-part SR item from the core item list.

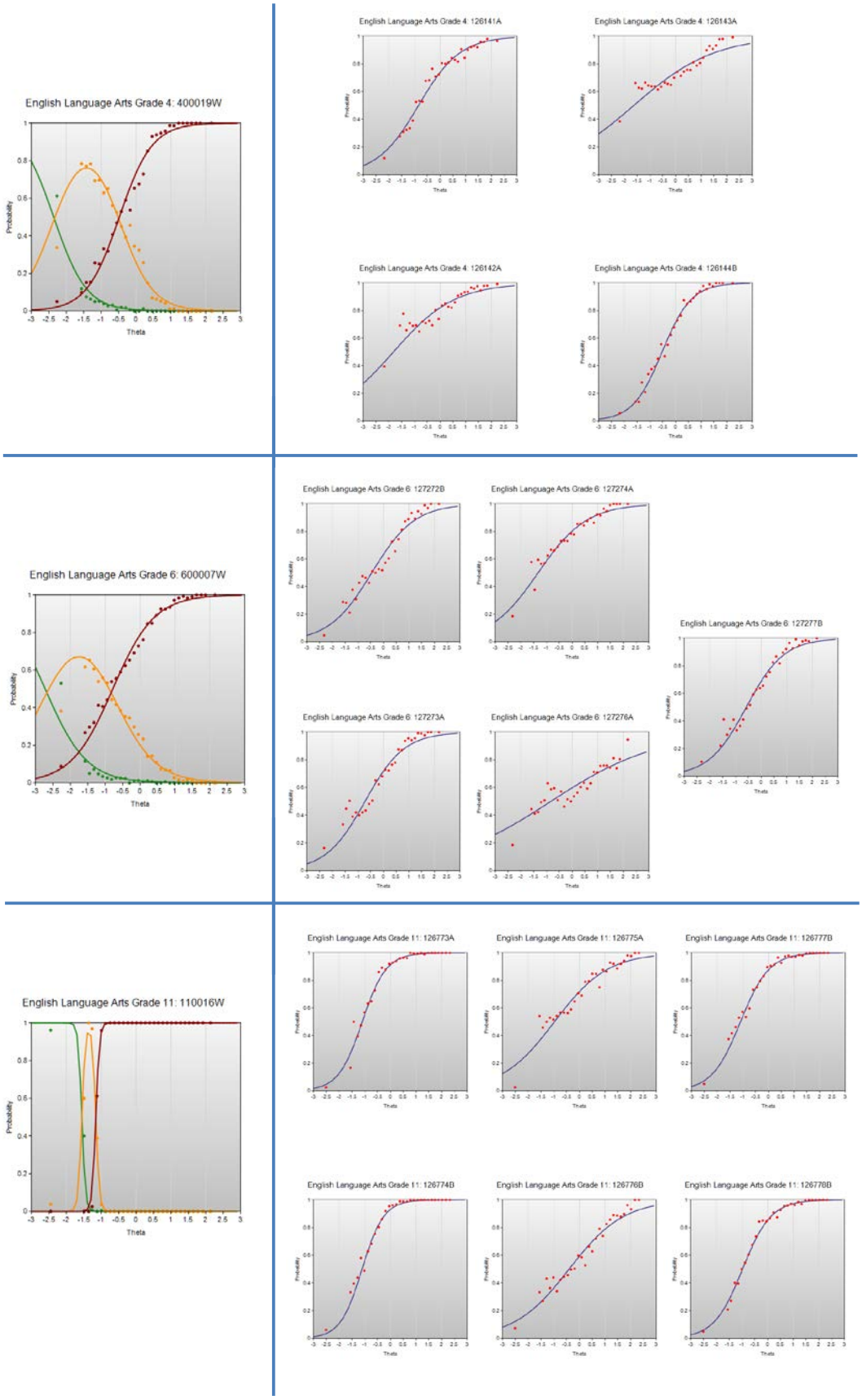
Tier 1 Writing Prompt Items

Tier 1 writing prompt (WP) items were a part of ELA tests for all grades. Four to six independently 0/1 scored selected response (SR) items were included in each Tier 1 WP item, where the number of SR items varied depending on the grade level. As we saw with the multi-part SR items, the answer to the previous item is always given to the students to minimize the local dependence of the items. When Pilot 2 was administered, Tier 1 WP items were scored using a “combined” approach as follows:

- First, sum all items included in a particular Tier 1 WP item.
- Then, a "0" was assigned to a summed scored of 0, a "1" was assigned to a summed score of 1 or 2, and a "2" was assigned to a summed score of 3 or greater, resulting in a final score of 0, 1, or 2.
- The scoring rule was applied to all Tier 1 WP items regardless of the number of items the item consisted of.

Figure 3 provides an example of item-level fit plots for three Tier 1 WP items: one from each of grades 4, 6, and 11. The plots on the left column demonstrate the item fit for the “combined” approach where a polytomous IRT model (i.e., graded response model [GRM]) was fitted to the data. On the right column, fit plots produced by fitting the 2PL model to each SR item are presented. The number of items included in each Tier 1 WP item is different in each case.

Figure 3. Example Item-Level Fit Plots for Tier 1 Writing Prompt Items



The fit plots for the “combined” approach indicate that the items are quite easy, which can be readily seen from the fact that the model-predicted curves are shifted to the left side of the plot. The classical statistics also show that these “combined” items are very easy (*p*-values were .88, .80, .80, .84, .90, .94, and .93 for grades 3, 4, 5, 6, 7, 8, and 11, respectively.). It seems that the *p*-values are falsely inflated by combing the scores and assigning new score values, especially for Tier 1 WP items that consist of six SR items. Students only have to answer half of those items correctly to obtain a maximum score on the Tier 1 WP item because students with summed scores of 3 and above are assigned the maximum final score of “2.” Also, the model-data fit for the “combined” items were problematic. Especially for the grade 11 Tier 1 WP item, the estimated *a*-parameter equaled 11.34, an extremely abnormal value, with a huge standard error of .61, indicating that the model does not adequately fit the data.

The “separate” approach shows that items included in a particular Tier 1 WP item can be very different in their model-data fit and shape of item response functions (IRFs), and we can more accurately estimate the shape when we fit the 2PL model to each separate item. By combining the items into a single polytomously scored item, we lose the opportunity to individually evaluate each item, and to make different decisions depending on the IRT characteristics of each item.

Recommendations

For the foundational items, multi-part selected response (SR) items, and Tier 1 writing prompt (WP) items, Measured Progress Psychometrics recommend 0/1 scoring for each item separately based on several reasons:

- TAC recommended in June, 2014 that NCSC should
    - continue to investigate scoring for foundational items; and
    - dichotomously score multi-part SR items, while also checking partial credit scoring;
    - focus on gathering high quality data before finalizing scoring of foundational and multi-part SR items.
- For these special item types, we believe that treating each item worth a point should enhance the model-data fit in terms of IRT.
    - For example, the 2PL model gives different discrimination parameters for each item, whereas a polytomous model would force a single discrimination parameter to be used.
- For multi-part SR items and Tier 1 WP items, the dependency among the multiple parts is removed by stating the correct answer to the previous item.
- We can gain additional psychometric information by treating items independently.
- Item statistics will correspond with each unique item, making item evaluation and documentation straightforward.
- Individual items could be revised or replaced without having to pull the entire set.
- If more or less weighting on these items is desired, we can apply the different weights independently of the statistical estimates of the items using appropriate psychometric models, when we calculate student scores.