PAT Inquiry and Problem Solving in STEM Contexts

Technical Report March 2019

Ling Tan
Steven Kambouris
Clare Ozolins

Australian Council *for* Educational Research

CONTENTS

List of figures

List of tables

1 OVERVIEW

1.1 Background

PAT Inquiry and Problem Solving in STEM Contexts (also referred to as PAT STEM Contexts) is an assessment of students' knowledge and skills in contexts that allow for the inclusion of questions that address content descriptions from the Australian Curriculum domains of Science, Mathematics and Technologies. The PAT STEM Contexts assessment has been developed in recognition of the increasing interest in, and importance of, the development and assessment of students' knowledge and skills across Science, Technology, Engineering and Mathematics in an integrated manner (O'Connor, 2018).

The PAT STEM Contexts trial tests were delivered online to allow for the inclusion of dynamic stimuli and interactive response formats. The format of the assessments is appropriate for the type of questions being asked and skills being assessed. Each item is aligned to a content description of the Australian Curriculum, drawn from the domains of Science, Mathematics, and Technologies. These are the domains that contain learning outcomes relevant to 'STEM'. The majority of PAT STEM Contexts items are also assessed in real-world contexts, rather than addressing facts, knowledge and skills in isolation. In many cases, the contexts selected allow for questions to be asked from more than one of these domains. Some units and their associated questions retain an emphasis on Science content descriptions across all three strands, allowing for inclusion of learning to be assessed across Science Understanding; Science as a Human Endeavour; and Science Inquiry Skills strands. Each item is also linked to one of three cognitive skills, Knowing, Applying, and Reasoning, which are based on the classifications used in the Trends in International Mathematics and Science Study (TIMMS) 2019 assessment frameworks for Science and Mathematics (Mullis, I. V. S., & Martin, M. O., 2017).

Some stimulus materials for real-world contexts include animations, allowing for phenomena and processes to be presented in a dynamic manner. Students interact with items in a variety of ways; some items involve the use of drag-and-drop and hotspot functionality, in addition to items using multiple-choice and complex multiple-choice formats. There are also some items that require students to enter a numerical response (cloze item format). Some items are reported as fully correct (2 score points), or partially correct (1 score point), providing further diagnostic information.

1.2 Trial test design

The trial material was focused on the upper years of primary school (Years 5 and 6), with approximately 50 per cent of items addressing these two year levels. Some materials also targeted either Years 3 and 4, or the first two years of secondary schooling, Years 7 and 8. Some items were trialled at two year levels to collect empirical evidence about which year level each item was best suited to. By including the same items at two year levels, we can compare item statistics and the performance of students at the two year levels on these items. For this reason, the same set of items was trialled with both Years 5 and 6, and another set of items was trialled with both Years 7 and 8. In total, eight trial test forms were developed for the field trial: one trial test form for Year 3 students, two test forms for students in Year 4, three test forms for Years 5 and 6, and two test forms for Years 7 and 8.

Each test form contained a mixture of multiple-choice items, interactive items and/or cloze items. Test form lengths ranged from 22 to 33 items. Table 1 shows the total number of items in each test form and the number of items by format. In each trial test form, the majority of items (70–85%) were multiple-choice questions, including both simple multiple-choice questions and complex multiple-choice questions. The interactive items (drag-and-drop and hotspot items) account for 12% to 23% of items in a test. Only a small number of cloze items were trialled in test forms at Years 6 to 8 (up to 9% in each test).

Table 1 Trial test forms

Test form	Item format					Number of items
	Simple multiple-choice	Complex multiple-choice	Drag-and-drop	Hotspot	Cloze	
3	12	5	4	1		22
4A	20	6	4	1		31
4B	20	4	6	1		31
5A or 6A	22	5	3	2	1	33
5B or 6B	22	5	4	1	1	33
5C or 6C	19	7	2	2	3	33
7A or 8A	16	5	3	4	2	30
7B or 8B	22	2	4	2		30

1.3 Description of trial samples

The PAT STEM Contexts trial was conducted in Australian schools in 2017. Test forms were administered online, delivered via ACER's online assessment and reporting system (OARS). Schools using ACER's other online PAT assessments could opt for their students to participate in the PAT STEM Contexts test trial. The characteristics of the trial samples are detailed in Section 2 of this report.

1.4 Trial analysis

After the trial administration, student responses were analysed to assess the psychometric properties of all trial items and tests. During the initial analysis, the component responses in complex multiple-choice (CMC) items were split into separate responses and could then be treated as responses to separate items. This provided test developers an opportunity to diagnose the function of each part of the CMC item. After this initial analysis, responses to CMC items were scored as single items and analysed together with other non-CMC items.

Responses from each trial form were analysed separately using the Rasch model. These analyses indicated how well the items in each form fitted the Rasch measurement model and revealed items that did not perform as well as expected. A total of 97 items were trialled. Of these, 13 items (about 13%) were judged to have unsatisfactory psychometric properties and were deleted from the pool available for constructing the final test forms. The remaining items functioned well statistically, and were well targeted for difficulty as described by the test construct and assessment framework.

A common item equating design was used to equate tests across year levels onto a single scale. This design made it possible to locate all items in the trial forms on a new scale, referred to as the PAT STEM Contexts scale (patstem). This meant that student performance results from different trial forms were directly comparable. Details of the equating design and equating results are provided in Section 7 of this report.

During test development, a goal was to avoid items that might favour one subgroup of students over another, for example girls compared with boys. Differential item functioning (DIF) analysis on gender was performed on all trial items. Any item exhibiting a statistically significant difference in subgroup performance for students of the same ability was flagged and subject to content analysis by test developers. Any items with content or context bias would potentially be excluded from the final assessment forms. Trial item analysis is described in more detail in Section 5 of this report.

1.5 Cut scores

The PAT STEM Contexts scale has been categorised into five levels, or bands, of achievement. Each level qualitatively describes the skills and understandings a student has demonstrated based on their performance on a PAT STEM Contexts test. The levels of achievement are independent of the tests. They can be used to compare student results obtained from different PAT STEM Contexts tests and assessed at different times. The determination of cut scores is described in the Section 8 of this report.

2 DESCRIPTION OF TRIAL SAMPLES

This section covers the demographic characteristics of students who participated in the online trial in 2017. A total of 7273 students from 63 schools across Australia participated in the PAT STEM Contexts trial. Table 2 shows the number of students who participated in the trial by test form. The number of participants was smallest in trial form 6B (only 296 students), and the highest in trial form 3 (1347 students).

Table 2 Number of students by year level and trial test form

Year level	Test forms								Total
	3	4A	4B	5A or 6A	5B or 6B	5C or 6C	7A or 8A	7B or 8B	
3	1347								1347
4		463	800						1263
5				431	356	723			1510
6				671	296	612			1579
7							397	350	747
8							404	423	827
Total	1347	463	800	1102	652	1335	801	773	7273

Overall, about 53% of the students were female. The proportions of male students among the participants at secondary school level were low (25% in Year 7 and 21% in Year 8). Table 3 shows the proportion of students by gender at each year level. There is a high proportion of missing gender information (24%) at Year 8. These students did not specify their gender information.

Table 3 Proportion of students by year level and gender

Year level	Number of students	Female (%)	Male (%)	Unspecified (%)
3	1347	50	50	
4	1263	50	50	
5	1510	52	48	
6	1579	48	52	
7	747	75	25	
8	827	56	21	24
Total	7273	53	44	3

Students from 63 schools across Australia participated the PAT STEM Contexts trial. No schools from the Northern Territory participated the trial.

Table 4 Number of students by year level and state/territory

Year level	NSW	VIC	QLD	WA	SA	ACT	Total
3	136	613	250	301	47		1347
4	181	532	179	332	39		1263
5	176	640	225	429	40		1510
6	119	665	308	436	51		1579
7	35	495	163		51	3	747
8	23	672	87		44	1	827
Total	670	3617	1212	1498	272	4	7273

3 DATA CLEANING AND PRE-PROCESSING

3.1 Data cleaning

Prior to analysis, item response data were checked for unexpected or invalid values. For example, valid codes of 'A', 'B', 'C' or 'D' were expected from the responses of simple multiple-choice items. Response lengths and valid codes were also checked for complex multiple-choice items, hotspot items, drag-and-drop items. The entered responses to short response items were checked for validity by test developers.

Item keys for simple multiple-choice items were checked for anomalies using item analysis statistics produced using ACER ConQuest software (Adams, Wu and Wilson, 2015). For example, point biserial for each option for each item was checked to see if the correct responses had the highest positive correlation with the total scores of the rest of the items in a test.

3.2 Handling of missing data

Students may leave items unanswered either because an item was too difficult for the student, or because the student ran out of time and so did not attempt it. In the former case, the student has seen the item and chosen not to provide a response. In the latter case, the student did not see the item at all. These two types of omitted or missing data are coded differently in the ACER online testing system.

If missing responses where students did not see the item are treated as incorrect responses, item difficulties may be overestimated. To avoid this, omitted responses on the items that were not seen by the students were treated as non-administered in item difficulty estimation. Both types of missing responses are considered incorrect for the purpose of estimating student achievement scores.

3.3 Splitting of complex multiple-choice data

During the initial item analysis, the component responses of each complex multiple-choice (CMC) item were split into separate responses as if from multiple items. This offered test developers opportunities to diagnose CMC item component performance and to determine the scoring rules for CMC items based on the empirical data.

3.4 Scoring of short response items

The frequencies of responses to short response items (cloze items) were tabulated and provided to test developers to review and modify scoring rules if necessary based on the trial response data. The updated scoring rules were then used for scoring the short response items for use in the item calibration data.

4 SCALING METHODOLOGY

Response data from the PAT STEM Contexts trial tests were fitted to the partial credit Rasch measurement model (Rasch, 1980; Masters, 1982). This model is expressed mathematically as:

$$P(x_i|\theta_n) = \frac{\exp\sum_{j=0}^{x}(\theta_n - \delta_i + \tau_{ij})}{\sum_{h=0}^{m_i}\exp\sum_{j=0}^{h}(\theta_n - \delta_i + \tau_{ij})} \quad x_i = 0,1,\ldots,m_i$$

where $P(x_i|\theta_n)$ is the probability of person *n* to score *x* on item *i*. θ_n denotes the person's level of the latent trait, the item parameter δ_i gives the location of the item on the latent continuum, and τ_{ij} denotes *j*th step parameter of item *i*.

The Rasch model assumes that the achievement of a student can be captured with a person parameter and the difficulty of an item can be captured with an item parameter. The model allows student achievement and item difficulty to be displayed on the same scale. This is because they are both measurements of the same construct: student achievement reflects the level of skill and understanding demonstrated by the student; item difficulty reflects the level of skill and understanding required to answer the item correctly. The high-achieving students and difficult items are located higher on the scale than low-achieving students and easy items.

PAT STEM Contexts scale scores can be used to directly compare student performance on two separate occasions, even if different test forms are used. This is possible because observed raw test scores on any PAT STEM Contexts test can be converted to locations on the PAT STEM Contexts scale. It is not meaningful to compare observed raw test scores from different test forms (even if they are expressed as percentages), because observed raw test scores and percentages do not take into account the relative difficulty of the tests.

The PAT STEM Contexts scale is an interval scale: a change of one unit corresponds to the same amount of change in achievement at all locations along the scale. The measurement scale has no upper or lower limits. When data are fitted to the Rasch model, the locations of items on the scale reflects their difficulty relative to other items and are independent of distribution of student achievement along the scale. Levels of achievement along the measurement scale can be qualitatively described, allowing the result of the assessment of a student to be reported in descriptive terms.

The Rasch model requires all items to have the same discrimination (while allowing for a degree of random variation) in order to have the property of invariant item ordering. Other scaling methodologies based on Item Response Theory (IRT) models include the two-parameter logistic (2PL) model and the three-parameter logistic (3PL) models. These aim to specify models that fit the observed data (rather than have data that fit the model as in Rasch) and introduce additional item parameters to achieve a better fit to the observed data. The 2PL model can provide a better fit to responses on items that may be unequally discriminating. The 3PL model provides a representation of student test-taking behaviour, for example guessing, and provides a better fit to the response data of multiple-choice items with the introduction of a guessing parameter. Wright (1999) argued that the crossing item characteristic curves in 2PL and 3PL cause the hierarchy of relative item difficulty to change at every ability level. This may constitute a threat to the construct validation of the instrument. In addition, items that vary in discrimination may be contaminated by item bias or may introduce extra dimensions (Masters, 1988; Wright 1992).

The Rasch model has the advantage that ability estimates have a one-to-one correspondence with the number-correct or raw test score. For the 2PL and 3PL models, an individual's ability estimate is based on their particular response pattern. Test scores are weighted by item discrimination in the 2PL model. Different response patterns resulting in the same raw test score may not produce the same ability estimate. In other words, the ability estimate depends not only on how many items were answered correctly, but also which specific items were answered correctly. It is often difficult to explain to students and schools how students receiving the same number-correct score can receive different scale scores. Ability scores that only depend on number correct scores are often more acceptable to students when individual results are reported.

In the 3PL model, individual ability estimates are adjusted for guessing, irrespective of whether the student has guessed in the test. The estimation of item guessing parameters also depends on the particular cohort of students sitting the test, which makes the equating of tests over time based on fixed item parameters for common items unreliable. As the 3PL model has more parameters to estimate, it requires larger sample sizes and is potentially susceptible to large estimation errors.

The Rasch measurement modelling approach aims to have a test that collects data that fit the Rasch model. This is usually done during test development with vigorous test piloting and item selection processes. The Rasch model supports the construction of described proficiency/achievement scales that not only report to students how well they are doing, but can also relate their performance to what they can typically do at their achievement level.

ACER ConQuest (Adams, Wu, and Wilson, 2015) was the software used for Rasch scaling analysis. This software provides tools for the estimation of a variety of different item response models and regression models. It was used for item calibration, and for generating weighted likelihood estimates (WLE) for person estimates. The transformations of student ability scores from logits to scale scores are presented in Appendix 4.

5 TRIAL ITEM ANALYSIS

5.1 Item analysis

Initial analysis of the 97 unique items in the PAT STEM Contexts trial indicated that 13 items had poor fit to the model or inadequate discrimination. These items were removed from further analysis and consideration for the final assessment forms. Item statistics for the remaining 84 trial items are provided in Appendix 1, including item difficulty (logit, scale score and facility), item discrimination (item–rest correlation), item fit (weighted mean square, its confidence interval and T value) and the number of students attempting each item (number of data points).

A range of statistics are produced as part of the item analyses. Item facility and discrimination statistics are obtained from classical test analysis; the other statistics are obtained from item response theory analysis. The item characteristic curve (ICC) of each item is provided in Appendix 2.

The *item facility* statistic expresses the percentage of individuals who are successful in answering each question or item on a test. The mean item facility for trial items across year levels was 51.4.

The *item discrimination* expresses the correlation between the individual's score and the aggregate score on the set of items in the same test. The item discrimination index used throughout this report is item–rest correlation, in which the aggregate score excludes the score of the item under examination. The mean item–rest correlation for trial items was -.32.

One of the *item difficulty* statistics is expressed in units of 'logits' – a metric used to measure the test results across different test forms on the same scale. In the Rasch model, individuals and items are measured on the same scale, allowing fair judgements to be made about the relative difficulty of the items. Importantly, it also makes it possible to judge the relative proficiency of students, in spite of the fact that they have been administered different tests that may have had different levels of difficulty.

The item difficulty expressed in scale score is a transformation of item difficulty from logits to scale scores as described in Appendix 4.

The *item fit* statistics are a measure of the extent to which an item is contributing to the measurement of the characteristic of interest. In the case of the item weighted fit (weighted mean square), values near 1 are desirable. An item weighted fit value greater than 1 is often associated with a low discrimination index, and an item weighted fit value less than 1 is often associated with a high discrimination index. The mean of the weighted fit values for a scale is 1.0.

The *item characteristic* curve describes the relationship between probabilities of correct responses and differences between person ability and item difficulty. It can be shown that, when the observed item characteristic curve (ICC) is steeper than the expected ICC, the item fit mean-square value is less than 1. When the observed ICC is flatter than the expected ICC, the item fit mean-square value is greater than 1.

The last column in the Appendix 1 table lists the number of students who saw the question. The minimum number of observations for any one item in the trial was 458.

5.2 Differential item functioning

During item development, every effort is made to avoid producing items that might favour one subgroup of students over another. Despite this, a proportion of items may be flagged with potential differential item functioning (DIF) as part of the statistical analysis. Investigating the reasons for a particular item showing DIF between particular groups involves looking for an explanatory connection between actual characteristics of the item and assumed or posited characteristics of the groups.

Gender DIF analysis was performed on all trial items by year level. The mean item difficulty in each of the two independent sets of item difficulties was centred at zero to adjust for group difference in ability. Any item in a subgroup with fewer than 100 observations was removed from DIF analysis, because of small sample size.

Figure 1 shows the DIF plot for gender by year level. On each DIF plot, an item is represented by one point on the plot. A red diagonal line serves as the reference line, with confidence interval limits indicated by the thin curved lines on either side of the reference line. If the relative item difficulty for an item is not different between the two groups (ie after taking their overall performance on the test into account), the point representing the item should lie on or close to the reference line. The distance of a point from the reference line indicates the magnitude of any potential DIF. Any item that falls outside the two lines representing the confidence interval limits may warrant investigation for potential DIF.

From Figure 1 (page 11), it can be observed that for all year levels, the majority of items fall within the confidence interval limits or are close to the confidence interval limits. A few items are relatively far outside the confidence interval limits.

Table 5 lists gender DIF analysis items with a difference of 0.6 and above, items that were flagged for additional attention. The difference for each item is calculated as the difficulty for the female students minus the difficulty for the male students. The table shows that six items significantly favoured female students and eight items significantly favoured male students. Items showing DIF are investigated for unfair content and where this is found to exist the items are not selected for final tests. In practice, the DIF is often not content-related but rather performance-related; that is, the favoured subgroup is simply better at the skills being assessed, for a variety of reasons. After review, no trial items were removed for content bias.

Table 5 List of potential gender DIF items

Item label	Year level	Difference in item difficulties (logit)	Standardised difference in item difficulties	Chi-square	*p*-value	Gender favoured
ST170301	8	-1.24	-2.08	4.32	0.04	female students
ST170302	8	0.78	2.14	4.58	0.03	male students
ST170503	8	0.80	2.89	8.34	0.00	male students
ST171403	7	-0.75	-2.70	7.30	0.01	female students
ST171602	3	-0.76	-3.47	12.06	0.00	female students
ST171701	5	-0.64	-2.45	5.99	0.01	female students
	6	-0.76	-2.45	5.98	0.01	
	8	-1.04	-2.81	7.92	0.00	
ST171704	6	-0.62	-2.35	5.51	0.02	female students
ST171803	5	0.77	4.35	18.88	0.00	male students
ST171804	5	0.66	3.06	9.39	0.00	male students
ST172101	6	0.75	3.05	9.29	0.00	male students
ST172201	4	0.65	3.09	9.56	0.00	male students
ST172601	7	0.76	3.03	9.16	0.00	male students
ST172603	6	-0.81	-4.19	17.52	0.00	female students
	7	-0.91	-3.18	10.09	0.00	
ST172703	5	0.71	2.92	8.53	0.00	male students

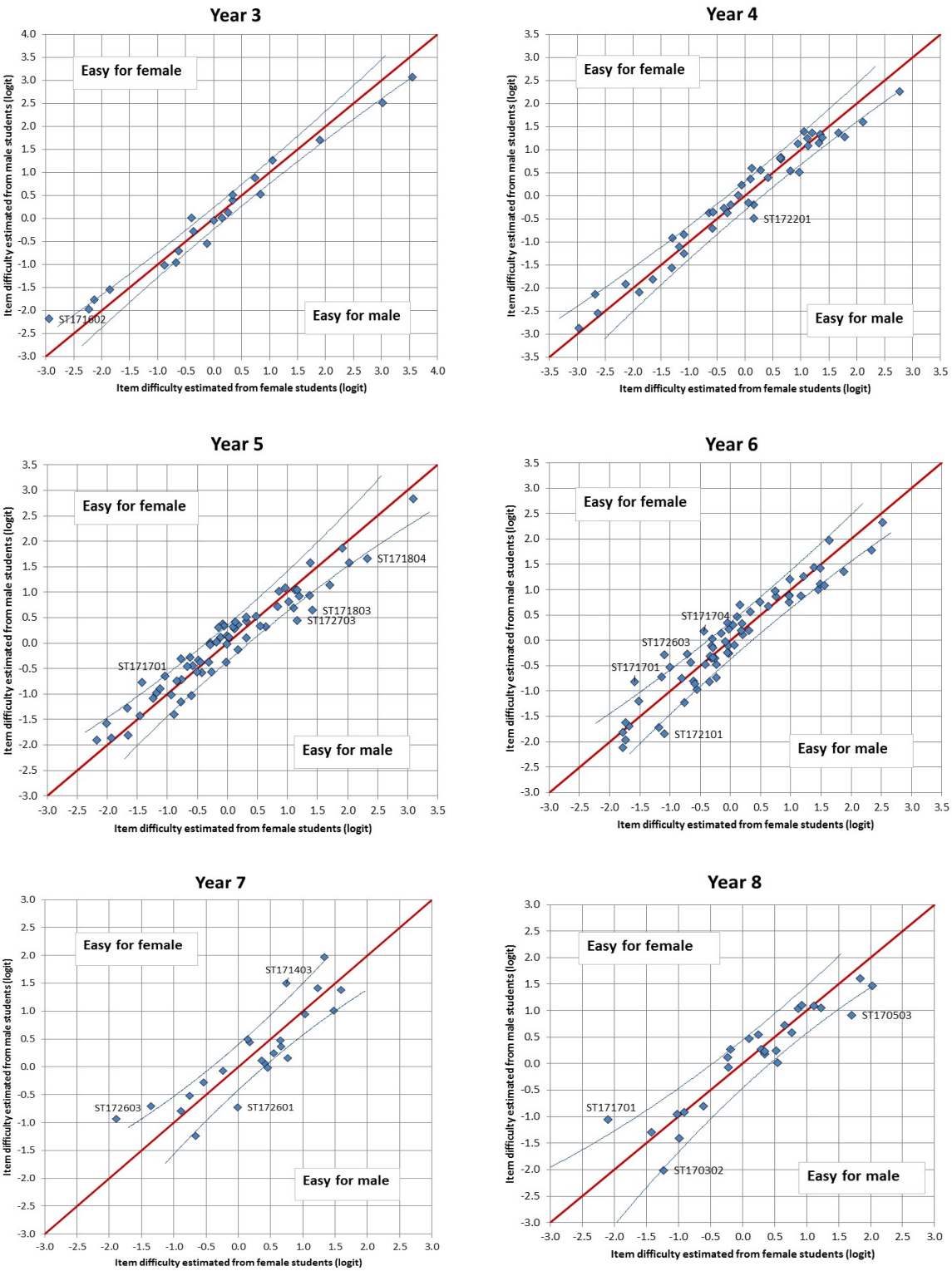


Figure 1 Gender DIF plots

6 TRIAL TEST ANALYSIS

6.1 Item–person maps

Using the Rasch model, individual person abilities and item difficulties are measured on the same scale. This makes it possible to judge the relative proficiency of students and relative difficulty of the items on the same metric. More importantly, an item–person map provides a visual indication of test targeting. A test is targeted to the trial samples if the test comprises items of varying difficulties and the distribution of item difficulty is aligned to the distribution of student ability. On each item–person map, the distribution of students is plotted on the left side of the map, and the distribution of items is plotted on the right side. The higher-ability students and more difficult items are positioned towards the top of the scale, and the lower-ability students and easier items are positioned towards the bottom. See Appendix 3 for item–person maps for Years 3–8.

6.2 Test reliability

Test reliability indicates the extent to which a test is consistent in measuring what it is intended to measure (in this case, inquiry and problem solving skills). Test reliability does not imply validity, but it is a necessary condition for validity. The test reliability coefficient is equal to the proportion of observed raw score variance that is attributable to true scores. Two test reliability indices were calculated for the PAT STEM Contexts trial tests: Cronbach's alpha and expected *a posteriori*/plausible value (EAP/PV) reliability, both shown in Table 6. The trial test reliabilities were close to or higher than 0.75 in all trial tests except for test form 8A (0.67). The overall Cronbach's alpha was not calculated due to a high proportion of missing data. This is denoted by asterisks (*) in Table 6.

Table 6 Trial test reliabilities

Year level	Test form	Cronbach's alpha coefficient	EAP/PV reliability
3	3	0.73	0.74
4	4A	0.81	0.83
	4B	0.81	0.81
	overall	*	0.82
5	5A	0.84	0.85
	5B	0.83	0.84
	5C	0.81	0.81
	overall	*	0.83
6	6A	0.85	0.86
	6B	0.83	0.82
	6C	0.83	0.79
	overall	*	0.83
7	7A	0.74	0.74
	7B	0.77	0.79
	overall	*	0.76
8	8A	0.67	0.68
	8B	0.78	0.81
	overall	*	0.76

6.3 Correlations among strands

The assessment items are categorised by the same cognitive skills used by TIMSS assessment framework: Knowing, Applying, and Reasoning. Table 7 shows the latent correlations between these strands. The latent correlations do not have the problem of attenuation caused by measurement error in discrete ability estimates. The value of a correlation can range from –1.00 (perfect negative correlation) through 0.00 (no correlation) to 1.00 (perfect positive correlation). All the correlations shown in Table 7 are significant at the 0.01 level of confidence.

The correlations among PAT STEM Contexts strands across all year levels were estimated by fitting a multi-dimensional latent regression model using a Monte Carlo method in ConQuest. For each strand, delta-centred item difficulty parameters were estimated by fitting a unidimensional measurement model regressed on test level. Then, the item difficulty estimates from all strands in an aspect were entered as anchored values in a multi-dimensional model regressed on test level. The italicised values in the table are the EAP/PV reliabilities for each strand.

Table 7 Correlations between strands

	Applying	Knowing	Reasoning
Applying	*-0.77*		
Knowing	-0.71	*-0.59*	
Reasoning	-0.84	-0.69	*-0.74*

Table 7 shows that there is a strong positive correlation (.84) between Applying and Reasoning strands. This is an expected result and indicates that students with high (or low) reasoning skills tended to have high (or low) applying skills. The correlation between Applying and Reasoning strands is higher than the correlation between Applying and Knowing or Knowing and Reasoning (.71 and .69 respectively), indicating that Knowing is a different strand from Reasoning and Applying. The EAP/PV reliabilities for Applying (.77) and Reasoning (.74) are higher than the reliability of Knowing (.59). This is because there are fewer Knowing trial items (only 14 items). All the correlations between strands within each aspect are close to or above .7, indicating a coherent relationship between strands as defined by the assessment construct.

6.4 Correlation with PAT Science

Link items from the established PAT Science assessment were included in each PAT STEM Contexts trial test form. The correlation between PAT STEM Contexts and PAT Science was estimated by fitting a multi-dimensional latent regression model using a Monte Carlo method in ConQuest. The correlation was found to be .81, indicating a reasonably strong relationship between these assessments. It is important to note that these PAT assessments are assessing different constructs and importantly, each of PAT STEM Contexts and PAT Science is reported on a separate scale.

7 TRIAL EQUATING DESIGN AND RESULTS

Two or three test forms were required at each year level in order to trial all the items developed. Separate tests trialled at the same year level were able to be equated via common items placed into each of the test forms. After equating, all trial items across test forms in a year level were concurrently calibrated. Figure 2 shows the common-item equating design among trial test forms in Year 4 (left panel), Years 5 and 6 (middle panel), and Years 7 and 8 (right panel). For example, the middle panel shows that Year 5 test forms A and B have shared items, and test forms B and C have shared items. The horizontal overlap between test forms indicates the proportion of items common to both test forms.

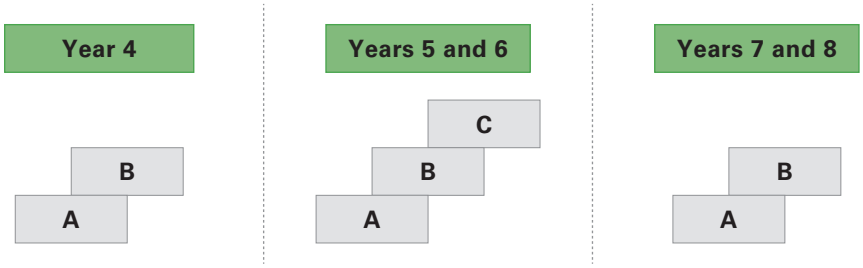


Figure 2 Test equating design within year levels

The PAT STEM Contexts trial items were trialled at Years 3–8. Some of these items were trialled at different year levels to understand the differences in item performance across year levels. For example, all items in Year 5 were also trialled at Year 6; and likewise for trial items in Years 7 and 8. Some trial items served as common items for the purpose of equating tests across year levels onto the same scale. The process of equating test forms across different year levels is known as vertical equating. For PAT STEM Contexts, vertical equating was achieved through the placement of common items in test forms between adjacent year levels. Figure 3 shows the vertical equating design of the trial tests. For example, about 50% of items in Year 7 tests were shared with Year 6 tests, and about 40% of Year 4 items were shared with Year 5 tests. All Year 3 items were shared with Year 4 tests.

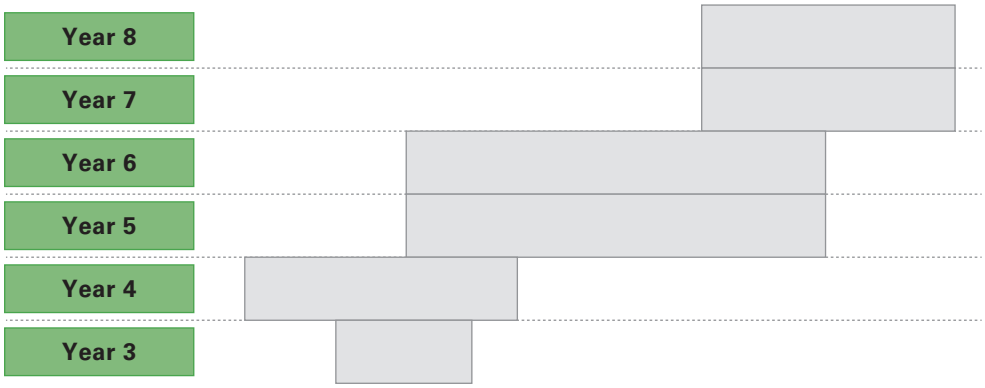


Figure 3 Test equating design across year levels

Common items between adjacent year levels were examined for the ordering of relative item difficulties in both year levels. This was to check whether common items between adjacent year levels were working as intended, and to confirm the validity of the vertical equating. The item parameters from the concurrent calibration obtained from each year level were used to conduct vertical equating.

Figures 4–8 (continued on page 16) show the scatter plots examining vertical equating by plotting the relative difficulties of common items between adjacent PAT STEM Contexts year levels. In each figure, the left panel shows the results before reviewing common items, and the right panel shows the results after excluding any misfitting items and the items with standardised difference greater than 3. The standardised difference is the difference of item difficulty estimates (adjusted for year level differences) divided by the pooled standard error. In each plot, the mean item difficulty in each of two sets of item difficulties was set to be the same to adjust for year level differences in ability. It can be observed that the vertically linked items were scattered around the diagonal identity line. The vertically linked items in each chart covered a wide range of item locations spanning at least 3.5 logits. The plots indicated that vertically linked common items were working well.

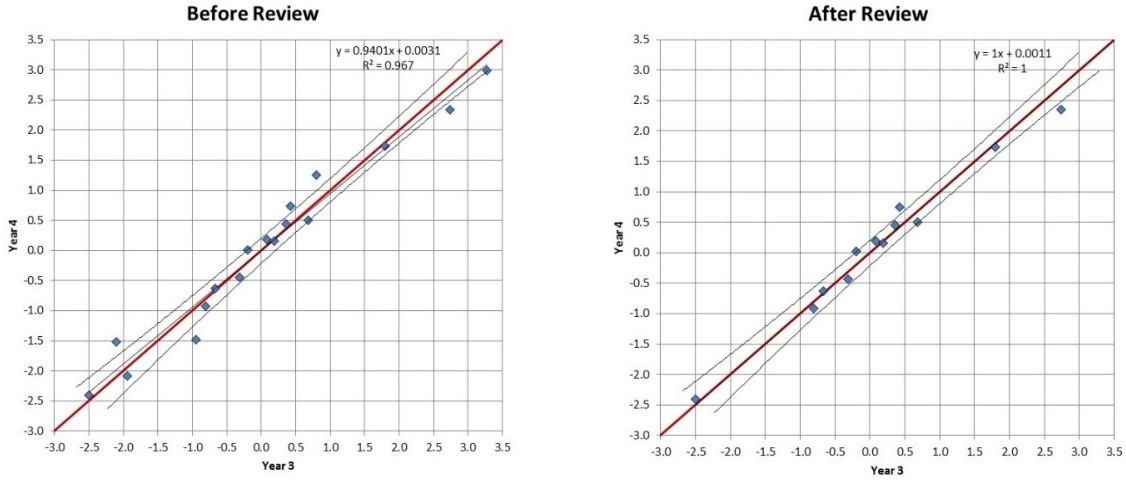


Figure 4 Vertical equating link item review for PAT STEM Contexts Year 3 and Year 4

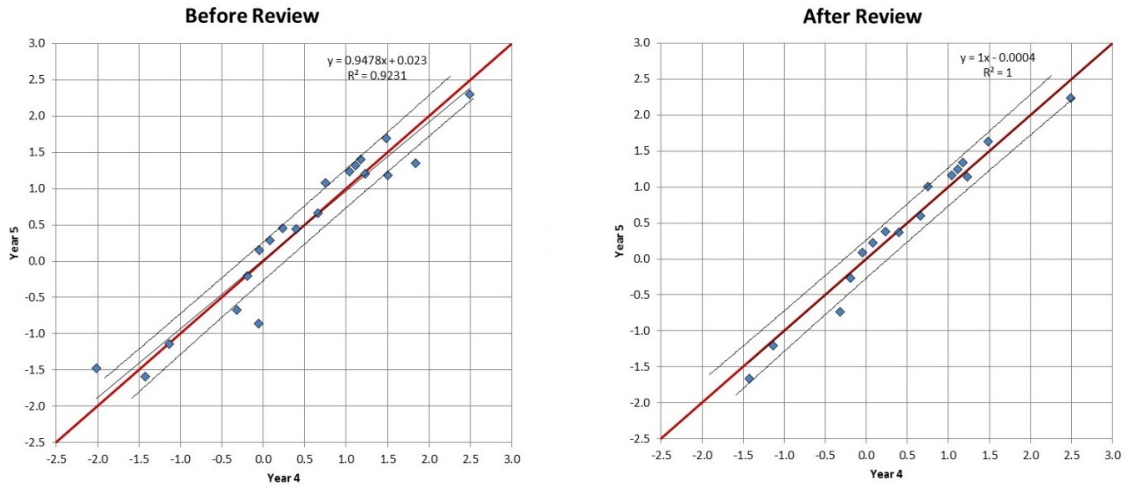


Figure 5 Vertical equating link item review for PAT STEM Contexts Year 4 and Year 5

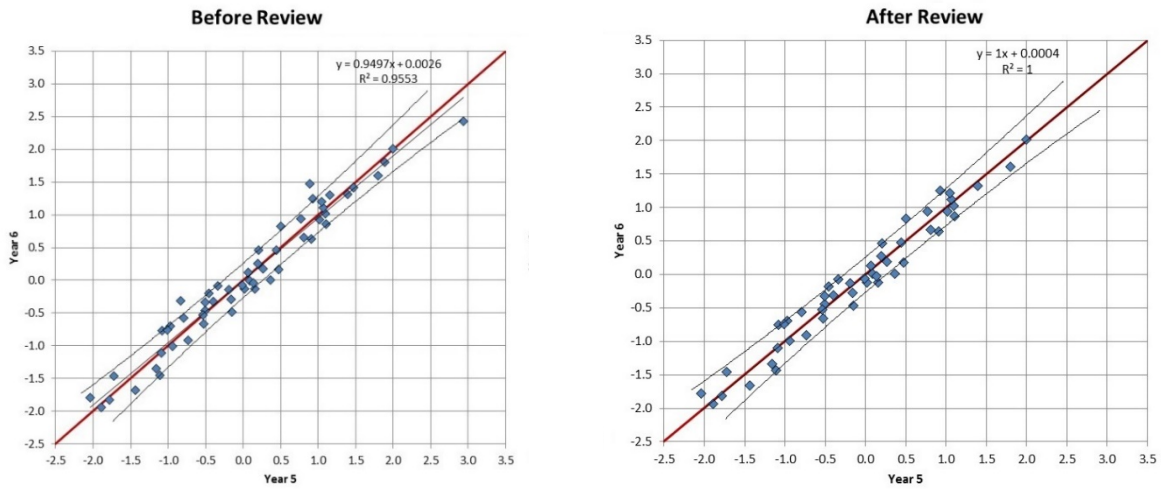


Figure 6 Vertical equating link item review for PAT STEM Contexts Year 5 and Year 6

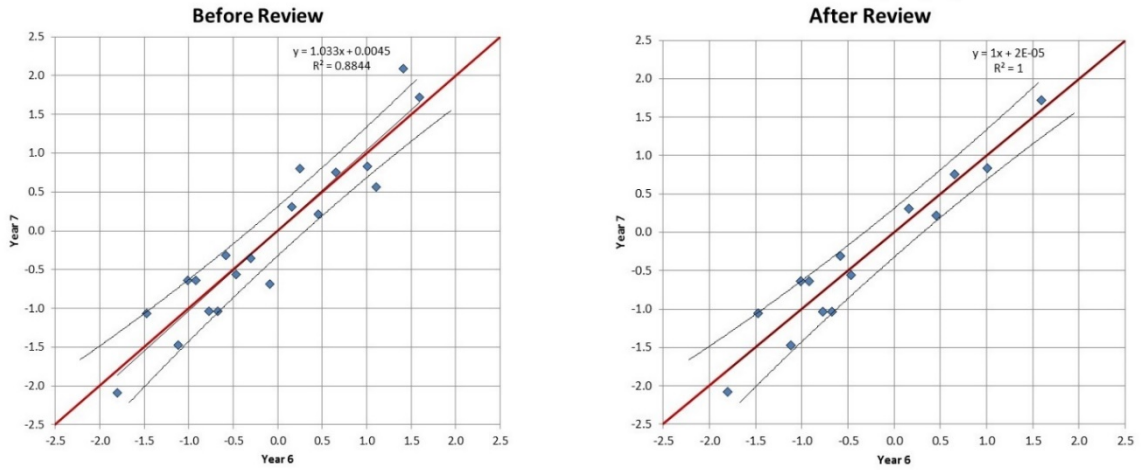


Figure 7 Vertical equating link item review for PAT STEM Contexts Year 6 and Year 7

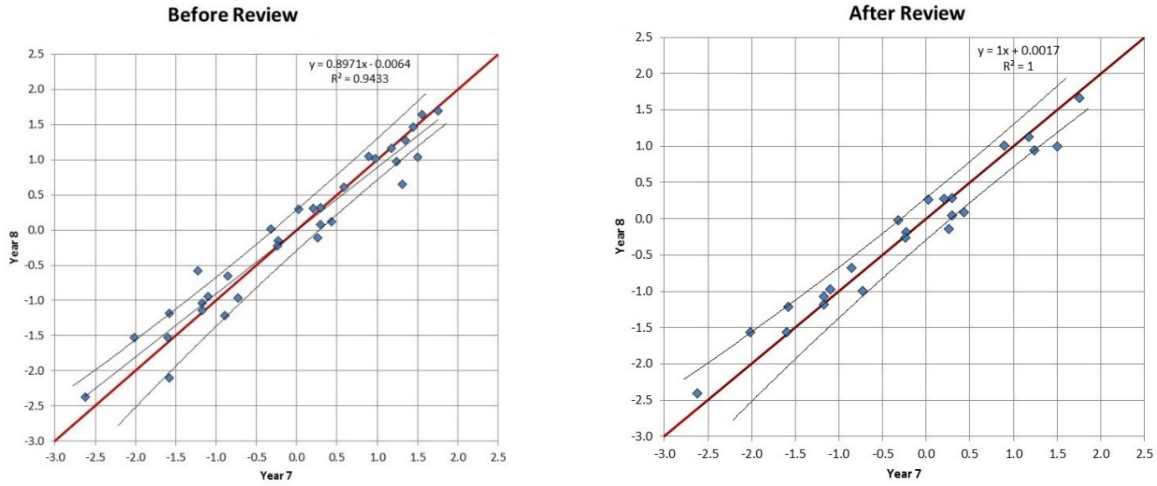


Figure 8 Vertical equating link item review for PAT STEM Contexts Year 7 and Year 8

8 SETTING CUT SCORES FOR ACHIEVEMENT BANDS

Five achievement bands were established for the PAT STEM Contexts assessments, based on the PAT STEM Contexts scale (patstem). Each band has a width of 10 scale score points. The cut scores defining the thresholds between adjacent bands are shown in Table 8, and the percentages of trial participants located within in each band by year level are shown in Table 9. Achievement bands are described in detail in a separate document, *PAT Inquiry and Problem Solving in STEM Contexts – Achievement band descriptions*.

Table 8 PAT STEM Contexts achievement band cut scores

Achievement band	Lower cut (patstem)	Upper cut (patstem)
5	≥135	
4	125	134
3	115	124
2	105	114
1		≤104

Table 9 Percentages of trial participants within each achievement band, by year level

Achievement band	Year 3	Year 4	Year 5	Year 6	Year 7	Year 8
5	1.7%	5.1%	7.5%	12.9%	15.7%	22.0%
4	10.4%	19.2%	23.7%	27.9%	36.2%	37.6%
3	28.9%	35.2%	36.4%	33.9%	34.6%	29.9%
2	35.3%	28.5%	24.3%	19.4%	12.0%	9.2%
1	23.7%	12.0%	8.0%	5.9%	1.6%	1.1%

9 PAT STEM CONTEXTS TEST FORMS

9.1 Test format

Following the detailed trial analysis, two PAT STEM Contexts test forms were constructed, with an initial emphasis on the middle years of schooling. The two forms are Middle Years A (recommended for Years 4, 5 and 6) and Middle Years B (recommended for Years 6, 7, 8). PAT STEM Contexts Middle Years A and Middle Years B have an emphasis on the Australian Curriculum outcomes at Years 5 and 6. Approximately half of the questions align to Years 5 and 6 outcomes. Middle Years A also contains questions aligned to Years 3 and 4, and Middle Years B also contains questions aligned to Year 7 and 8.

Based on the distribution of items aligned to the curriculum, and on the average difficulty of the tests, the Middle Years A assessment is most suitable for administration at Years 4, 5 and 6. The Middle Years B assessment is most suitable for administration at Years 6, 7 and 8. Each form begins with 9 practice items, followed by 34 assessment items. Students have one hour to complete the practice items and the test items.

The practice items have been designed to introduce test takers to the online testing interface and item formats that appear in each test form. The practice items are intentionally easy, so that students are not distracted by the content. They allow students to practise responding to multiple-choice items presented in a variety of formats, multiple-choice items presented as rows in tables that require a response for each row, dropping and dragging objects, selecting a hotspot, using the onscreen calculator, and to practise watching an animation, and scrolling down the page to see more content. Once students have completed the practice items, they are able to move on to complete the assessment items. Students can use the onscreen calculator provided as they complete the assessment.

9.2 Test difficulties

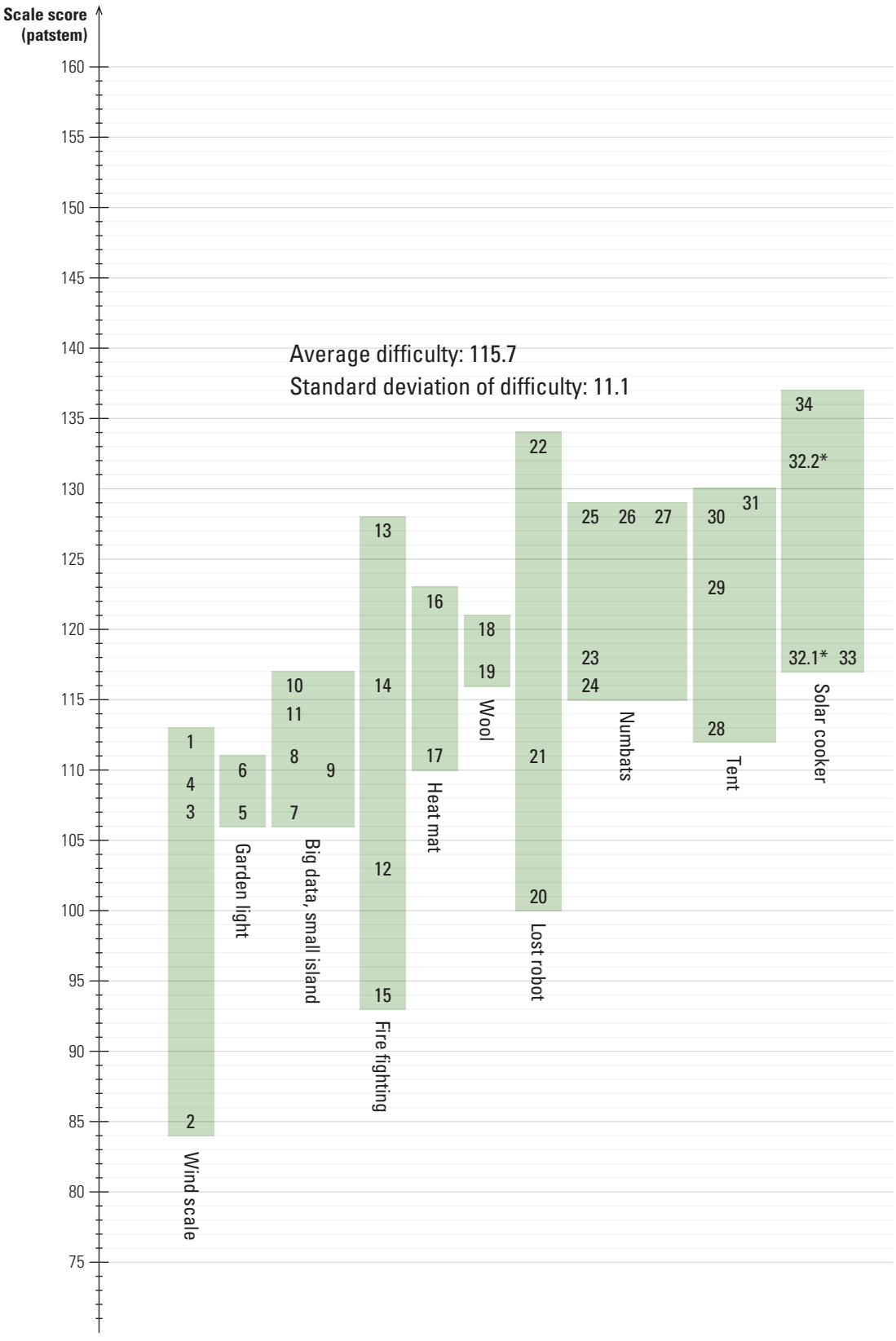
From the Rasch measurement analyses that were carried out, it is possible to report the mean difficulty of the items in each of the PAT STEM Contexts tests in scale score units. These mean item difficulties – or test difficulties – are shown together with their standard deviations in Table 10.

Table 10 PAT STEM Contexts test difficulties

PAT STEM Contexts tests	No. of items	Mean item difficulty	Standard deviation
Middle Years A	34	115.7	11.1
Middle Years B	34	124.1	10.7

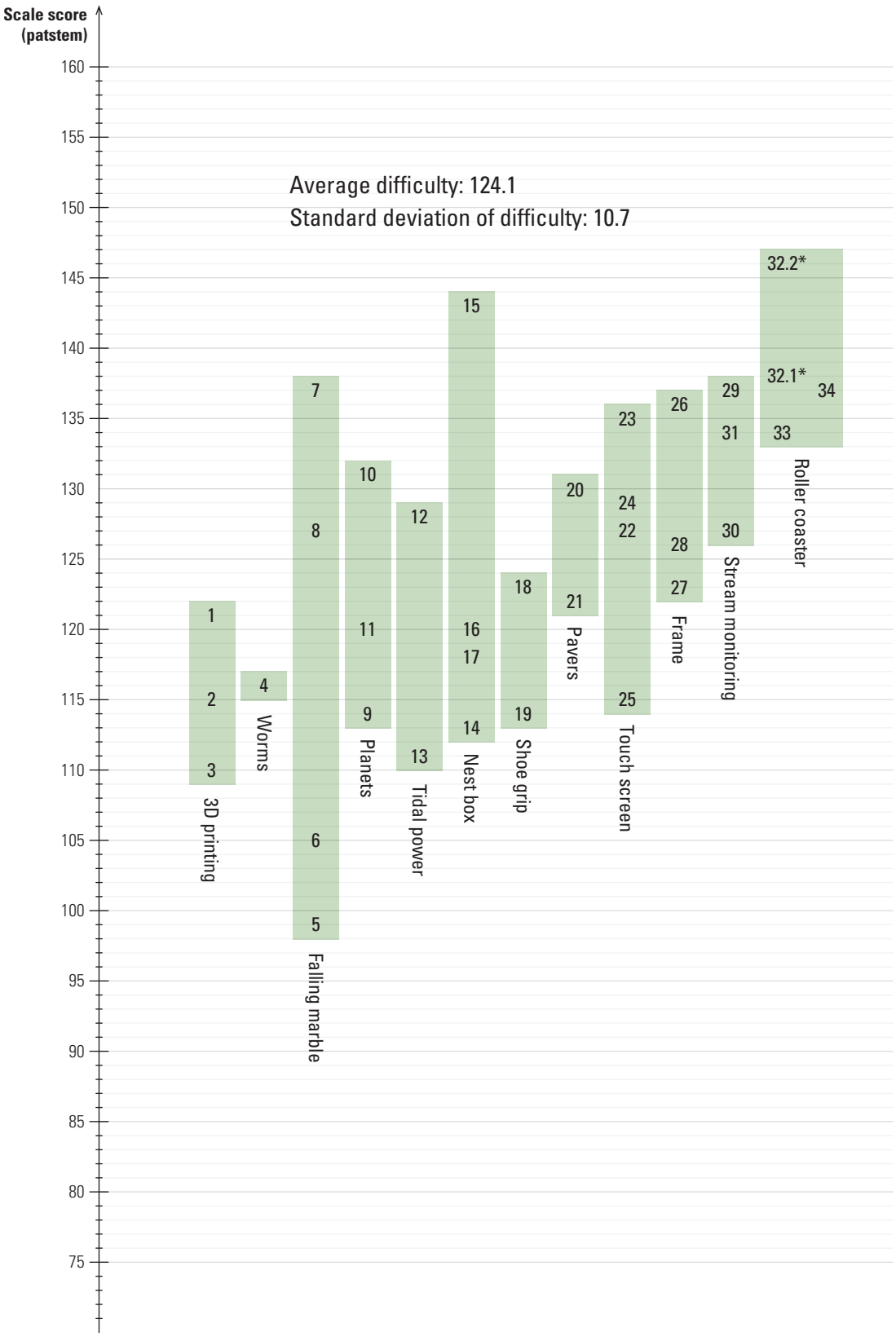
The locations of the PAT STEM Contexts items on the measurement scale are shown in Figure 9 and Figure 10 (pages 19 and 20). Items have been placed in shaded blocks according to the unit to which they belong. Each block illustrates the range of item difficulties within a given contextualised unit, consisting of two or more test items. The item number labels within each block identify the items belonging to the unit. Overlap in the difficulty of items and units, as well as an overall progression in difficulty over the course of the tests, can be clearly seen.

Score equivalence tables for PAT STEM Contexts Middle Years A and Middle Years B were created, based on the delta-centred PAT STEM Contexts trial item parameters estimated from concurrent calibration. After checking vertical links, response data from different year levels can be combined into a single data file for the concurrent calibration. The concurrent calibration places all trial items from different year levels on the same scale simultaneously in a single calibration. ConQuest was used to create tables showing the equivalence between raw scores and ability estimates expressed in logit values. Next, the ability estimates were transformed to the PAT STEM Contexts scale using logit-to-scale score transformation parameters. The scale score transformation formula is shown in Appendix 4. The score equivalence tables for the PAT STEM Contexts tests are shown in Appendix 5.



*Item 32 is a partial credit item.
32.2 represents the difficulty of scoring 2/2 on this item.
32.1 represents the difficulty of scoring 1/2 on this item.

Figure 9 Item distribution of PAT STEM Contexts Middle Years A



*Item 32 is a partial credit item.
32.2 represents the difficulty of scoring 2/2 on this item.
32.1 represents the difficulty of scoring 1/2 on this item.

Figure 10 Item distribution of PAT STEM Contexts Middle Years B

REFERENCES

Adams, R.J, Wu, M.L, and Wilson, M.R. (2015). *ConQuest* 4. [computer program] Camberwell: Australian Council for Educational Research.

Masters, G.N. (1982). A Rasch model for partial credit scoring. *Psychometrika*, 47, 149–174.

Masters, G.N. (1988). Item discrimination: when more is worse. *Journal of Educational Measurement*, 25:1, 15–29.

Mullis, I. V. S., & Martin, M. O. (Eds). (2017). TIMSS 2019 Assessment Frameworks. Retrieved from Boston College, TIMSS & PIRLS International Study Center website: http://timssandpirls.bc.edu/timss2019/frameworks/

O'Connor, G. (2018). PAT STEM Contexts Achievement Band Descriptions, Australian Council for Educational Research.

O'Connor, G. (2018). Conceptualising an Assessment Framework for PAT STEM Contexts, Australian Council for Educational Research.

Rasch, G (1980). *Probabilistic Models for Some Intelligence and Attainment Tests*. Chicago: MESA Press (original work published 1960).

Wright, B.D. (1992). IRT in the 1990s: Which Models Work Best? 3PL or Rasch? *Rasch Measurement Transactions*, 6:1, 196–200.

Wright, B.D. (1999). Fundamental measurement for psychology. In S.E. Embretson and S.L. Hershberger (Eds), *The New Rules of Measurement* (pp 65–104). Mahwah NJ: Lawrence Erlbaum Associates.

APPENDICES

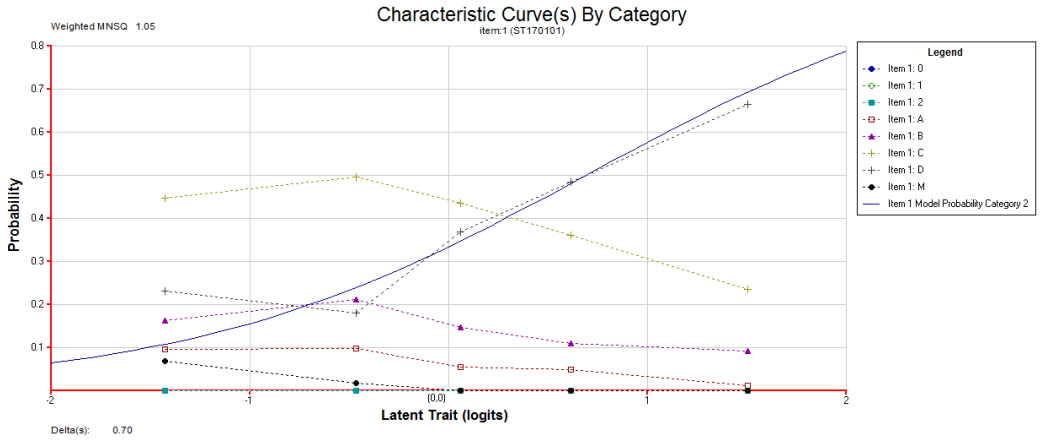
Appendix 1 PAT STEM Contexts trial item statistics

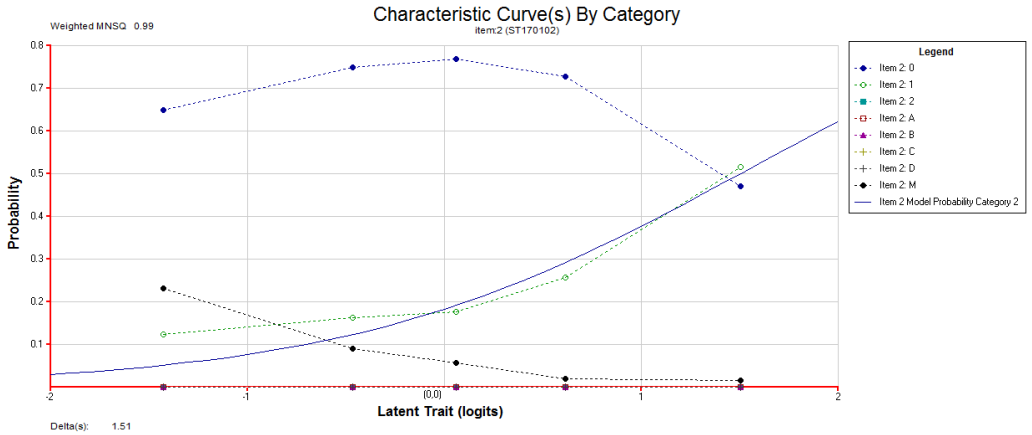
Item label	Cognitive skill	Item difficulty estimate (logit)	Item difficulty estimate (patstem scale score)	Facility	Discrimination (item–rest correlation)	Weighted fit		Number of data points
						Weighted mean square (95% confidence interval)	T Value	
ST170101	Applying	0.70	127	49.7	0.25	1.05 (0.97, 1.03)	2.6	1571
ST170102	Applying	1.51	135	32.9	0.31	0.99 (0.95, 1.05)	-0.6	1571
ST170103	Applying	0.94	129	41.5	0.19	1.08 (0.95, 1.05)	3.0	745
ST170104	Applying	-0.48	115	73.6	0.38	0.95 (0.94, 1.06)	-1.7	1538
ST170201	Reasoning	0.80	128	41.4	0.32	0.98 (0.96, 1.04)	-0.9	2048
ST170203	Applying	-0.87	111	75.0	0.27	1.02 (0.94, 1.06)	0.7	1979
ST170301	Applying	-2.09	99	89.3	0.37	0.90 (0.88, 1.12)	-1.6	1423
ST170302	Applying	-1.47	105	82.9	0.26	1.03 (0.91, 1.09)	0.7	1423
ST170303	Applying	1.73	137	24.5	0.21	1.11 (0.94, 1.06)	3.2	1423
ST170304	Reasoning	0.74	127	42.9	0.35	0.96 (0.96, 1.04)	-1.9	1423
ST170402	Knowing	1.37	134	36.0	0.30	0.99 (0.94, 1.06)	-0.2	798
ST170501	Applying	2.19	142	15.7	0.25	1.10 (0.87, 1.13)	1.5	773
ST170502	Reasoning	1.40	134	34.7	0.18	1.08 (0.94, 1.06)	2.6	773
ST170503	Applying	1.73	137	28.6	0.19	1.05 (0.93, 1.07)	1.3	773
ST170602	Applying	0.49	125	41.1	0.38	1.15 (0.94, 1.06)	5.0	1744
ST170603	Knowing	-0.24	118	57.1	0.29	1.07 (0.96, 1.04)	3.2	1725
ST170604	Applying	1.65	136	21.1	0.24	1.01 (0.93, 1.07)	0.4	1744
ST171701	Reasoning	-1.30	107	80.6	0.48	0.85 (0.92, 1.08)	-3.8	1423
ST171702	Knowing	-0.92	111	75.0	0.37	0.97 (0.93, 1.07)	-0.8	1423
ST171703	Reasoning	-1.03	110	76.7	0.44	0.90 (0.93, 1.07)	-3.0	1423
ST171704	Reasoning	-0.42	116	66.3	0.55	0.83 (0.95, 1.05)	-7.0	1423
ST171705	Reasoning	-0.63	114	70.1	0.40	0.94 (0.94, 1.06)	-2.2	1423
ST170901	Applying	-0.22	118	55.3	0.28	1.08 (0.97, 1.03)	4.1	2046
ST170902	Applying	-0.42	116	59.4	0.39	0.98 (0.96, 1.04)	-1.2	2046
ST170903	Applying	0.81	128	34.5	0.43	0.94 (0.96, 1.04)	-2.9	2046
ST170904	Reasoning	0.84	128	33.8	0.50	0.87 (0.96, 1.04)	-6.2	2046
ST170905	Reasoning	0.83	128	33.9	0.32	1.02 (0.96, 1.04)	1.2	2046
ST171001	Applying	-0.66	113	59.9	0.44	0.91 (0.94, 1.06)	-2.9	796
ST171002	Reasoning	0.28	123	40.3	0.26	1.06 (0.94, 1.06)	2.0	796
ST171003	Applying	0.84	128	29.4	0.26	1.04 (0.93, 1.07)	1.1	796
ST171004	Reasoning	0.91	129	28.3	0.28	1.01 (0.92, 1.08)	0.4	796
ST171901	Knowing	-2.95	90	88.2	0.25	0.99 (0.90, 1.10)	-0.1	1801
ST171902	Applying	0.20	122	39.5	0.27	1.06 (0.92, 1.08)	1.5	458
ST171101	Knowing	-0.35	116	46.2	0.30	1.01 (0.97, 1.03)	0.4	2139
ST171102	Applying	0.89	129	28.6	0.11	1.17 (0.92, 1.08)	4.2	796

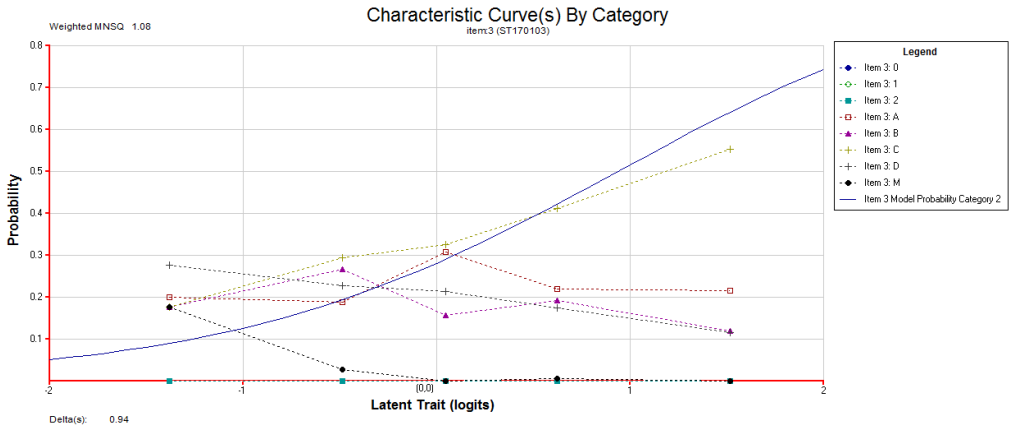
Item label	Cognitive skill	Item difficulty estimate (logit)	Item difficulty estimate (patstem scale score)	Facility	Discrimination (item–rest correlation)	Weighted fit		Number of data points
						Weighted mean square (95% confidence interval)	T Value	
ST171201	Applying	0.21	122	38.6	0.28	1.11 (0.97, 1.03)	6.2	2895
ST171203	Reasoning	-0.88	111	60.7	0.33	0.96 (0.97, 1.03)	-2.5	2895
ST171302	Applying	1.00	130	43.3	0.20	1.08 (0.96, 1.04)	4.1	1571
ST171303	Reasoning	0.19	122	61.2	0.30	1.01 (0.96, 1.04)	0.3	1505
ST171401	Applying	1.66	137	30.3	0.32	0.97 (0.93, 1.07)	-0.8	798
ST171402	Applying	0.67	127	50.6	0.30	1.01 (0.95, 1.05)	0.6	798
ST171403	Reasoning	1.39	134	35.6	0.21	1.07 (0.94, 1.06)	2.3	798
ST171501	Applying	-1.66	103	75.4	0.42	0.91 (0.96, 1.04)	-4.4	3691
ST171502	Reasoning	0.69	127	30.2	0.31	0.97 (0.97, 1.03)	-1.5	3691
ST171503	Reasoning	-0.35	116	50.8	0.28	1.07 (0.97, 1.03)	5.4	3691
ST171504	Reasoning	-2.59	94	87.4	0.44	0.91 (0.93, 1.07)	-2.6	3691
ST171601	Applying	-0.80	112	53.8	0.32	1.01 (0.96, 1.04)	0.8	1801
ST171602	Applying	-3.47	85	92.3	0.28	0.95 (0.86, 1.14)	-0.6	1801
ST171603	Knowing	-1.33	107	64.5	0.28	1.04 (0.96, 1.04)	1.8	1801
ST171605	Reasoning	-1.13	109	60.6	0.46	0.92 (0.96, 1.04)	-4.0	1801
ST171801	Reasoning	-1.87	101	79.3	0.29	0.99 (0.95, 1.05)	-0.3	3847
ST171802	Reasoning	-0.87	111	62.1	0.23	1.09 (0.97, 1.03)	5.7	3847
ST171803	Reasoning	1.32	133	20.9	0.33	0.93 (0.95, 1.05)	-3.0	3847
ST171804	Reasoning	2.03	140	12.6	0.21	1.00 (0.93, 1.07)	0.0	3744
ST172001	Applying	-2.07	99	78.7	0.33	0.94 (0.94, 1.06)	-1.9	2139
ST172002	Applying	0.77	128	30.7	0.38	0.96 (0.93, 1.07)	-1.0	796
ST172004	Reasoning	-0.03	120	46.6	0.31	1.03 (0.95, 1.05)	1.0	796
ST172101	Knowing	-1.33	107	78.2	0.34	0.98 (0.92, 1.08)	-0.6	1250
ST172102	Reasoning	-0.99	110	72.8	0.36	0.97 (0.94, 1.06)	-0.8	1250
ST172201	Knowing	-0.64	114	62.8	0.31	1.05 (0.95, 1.05)	2.0	1552
ST172202	Reasoning	1.07	131	28.7	0.30	1.02 (0.95, 1.05)	0.9	1552
ST172203	Reasoning	0.00	120	49.8	0.51	0.86 (0.96, 1.04)	-7.1	1552
ST172301	Knowing	1.01	130	29.8	0.16	1.14 (0.95, 1.05)	4.9	1552
ST172302	Applying	-0.16	118	53.1	0.36	1.00 (0.96, 1.04)	-0.1	1552
ST172401	Knowing	0.16	122	48.6	0.34	1.03 (0.96, 1.04)	1.4	1744
ST172402	Reasoning	-0.71	113	66.2	0.29	1.08 (0.95, 1.05)	3.2	1744
ST172502	Reasoning	1.11	131	30.5	0.05	1.22 (0.95, 1.05)	8.7	1900
ST172503	Reasoning	-0.43	116	61.8	0.34	1.01 (0.96, 1.04)	0.5	1900
ST172601	Applying	0.10	121	56.8	0.27	1.04 (0.96, 1.04)	2.1	1820
ST172602	Reasoning	-0.54	115	67.9	0.36	0.96 (0.95, 1.05)	-1.5	1892
ST172603	Reasoning	-1.02	110	76.2	0.40	0.90 (0.94, 1.06)	-3.5	1892
ST172701	Applying	1.59	136	26.8	0.23	1.04 (0.94, 1.06)	1.2	1423
ST172702	Reasoning	0.29	123	52.3	0.37	0.96 (0.96, 1.04)	-1.9	1423
ST172703	Applying	0.64	126	44.9	0.29	1.03 (0.96, 1.04)	1.6	1423

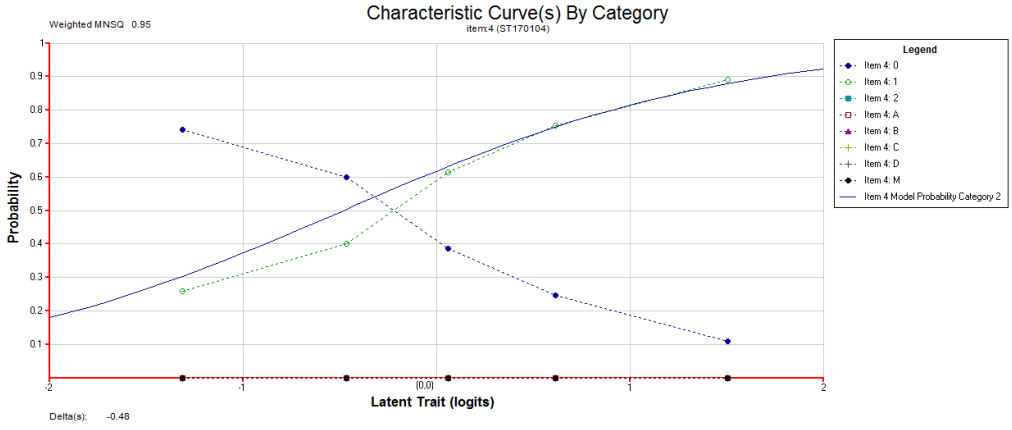
Item label	Cognitive skill	Item difficulty estimate (logit)	Item difficulty estimate (patstem scale score)	Facility	Discrimination (item–rest correlation)	Weighted fit		Number of data points
						Weighted mean square (95% confidence interval)	T Value	
ST172801	Applying	-0.01	120	53.2	0.41	0.96 (0.96, 1.04)	-2.5	1900
ST172802	Reasoning	-0.46	115	62.4	0.39	0.97 (0.96, 1.04)	-1.3	1900
ST172803	Applying	-0.27	117	59.2	0.43	0.94 (0.96, 1.04)	-3.1	1840
ST170701	Reasoning	-0.74	113	67.0	0.31	1.05 (0.94, 1.06)	1.5	1094
ST170702	Reasoning	2.33	143	13.2	0.34	0.91 (0.88, 1.12)	-1.4	1006
ST170703	Knowing	-0.01	120	52.5	0.34	1.04 (0.95, 1.05)	1.5	1094
ST170704	Applying	-0.19	118	56.1	0.36	1.02 (0.95, 1.05)	0.6	1094
ST170801	Knowing	1.88	139	26.3	0.13	1.09 (0.92, 1.08)	2.3	798
ST170803	Reasoning	0.26	123	52.5	0.32	1.04 (0.97, 1.03)	2.2	2048
ST170804	Reasoning	-0.64	114	70.3	0.32	1.03 (0.95, 1.05)	1.2	2048

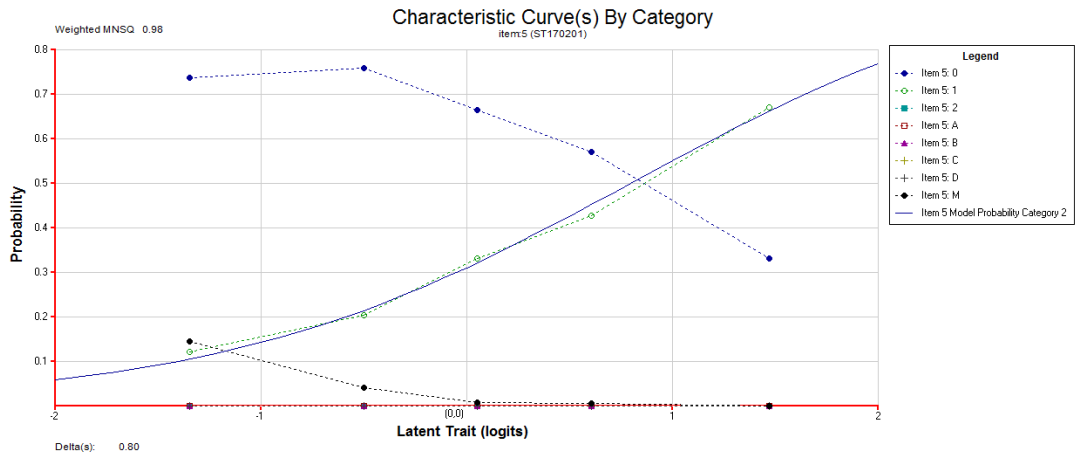
Appendix 2 Item characteristic curves (ICCs)

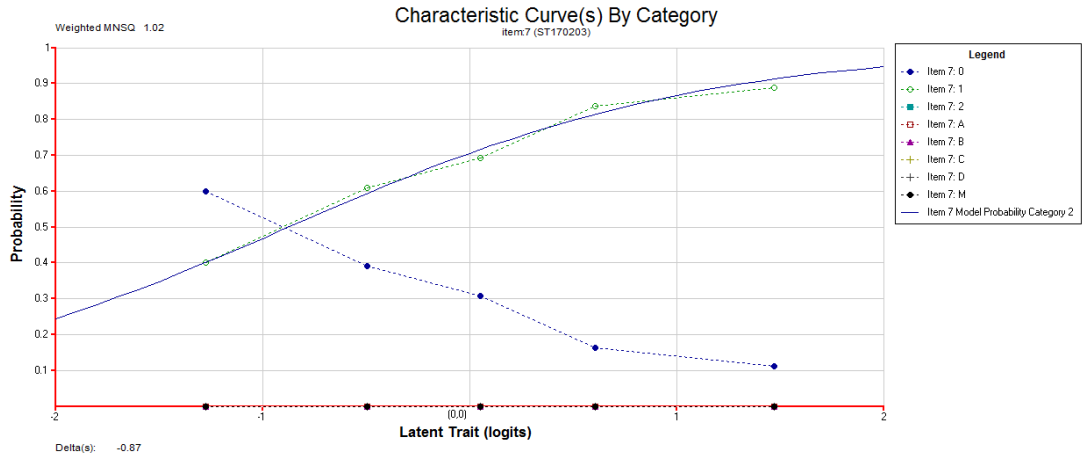


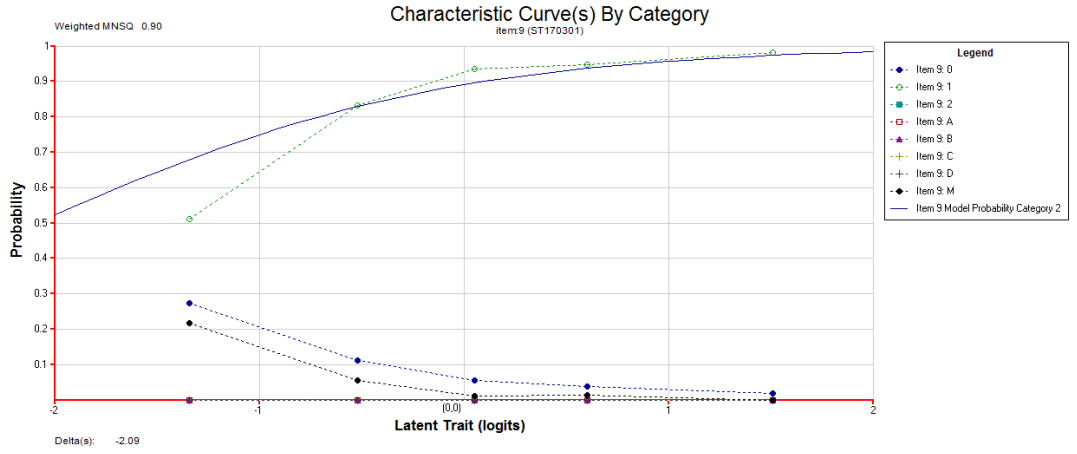


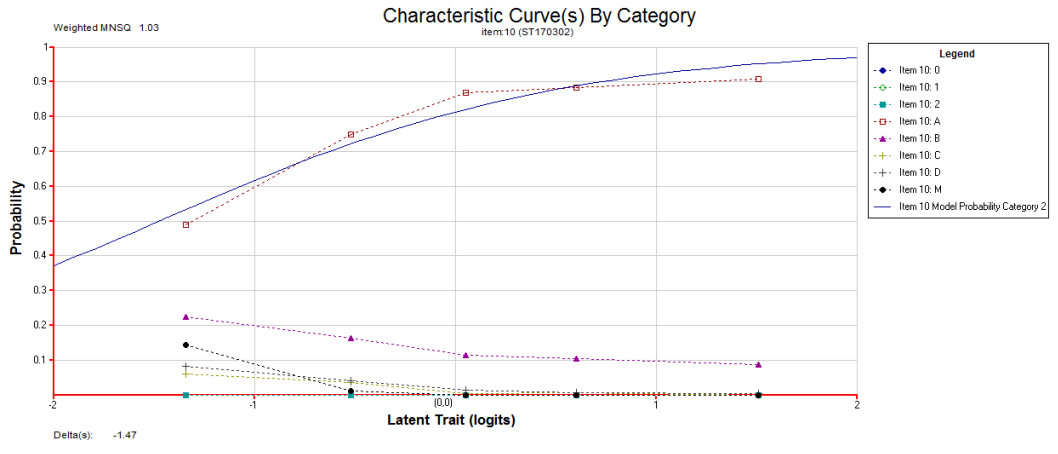


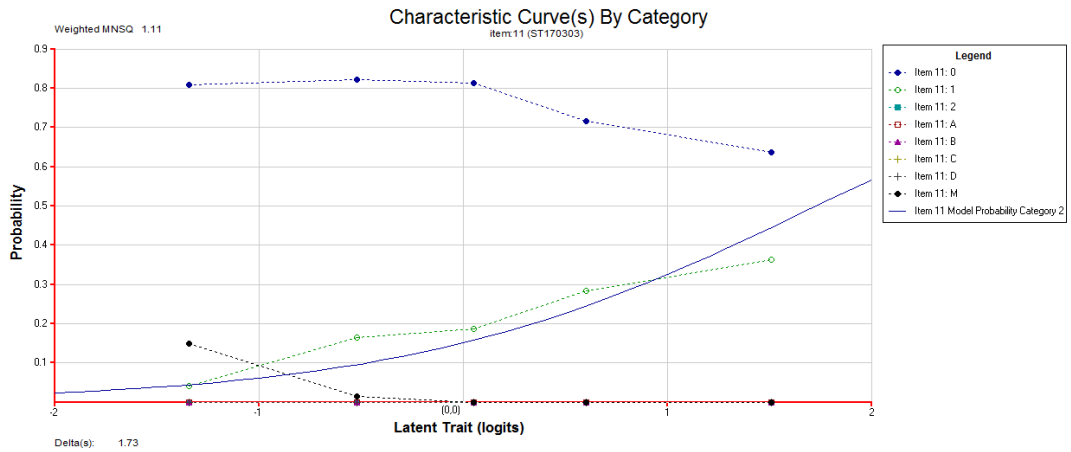


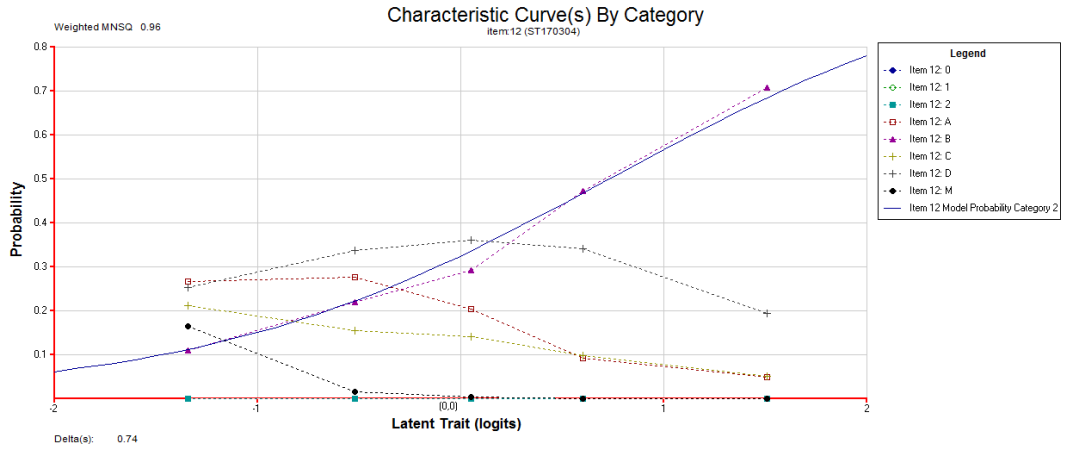


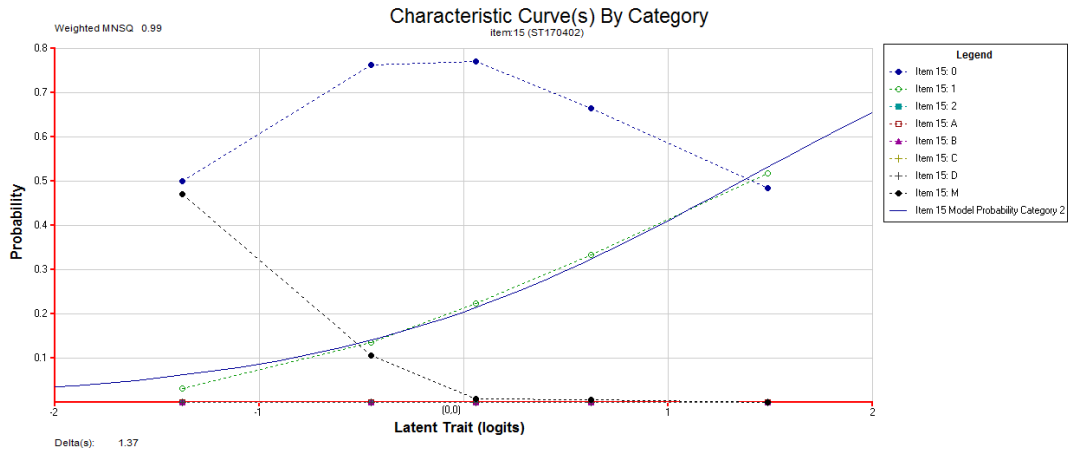


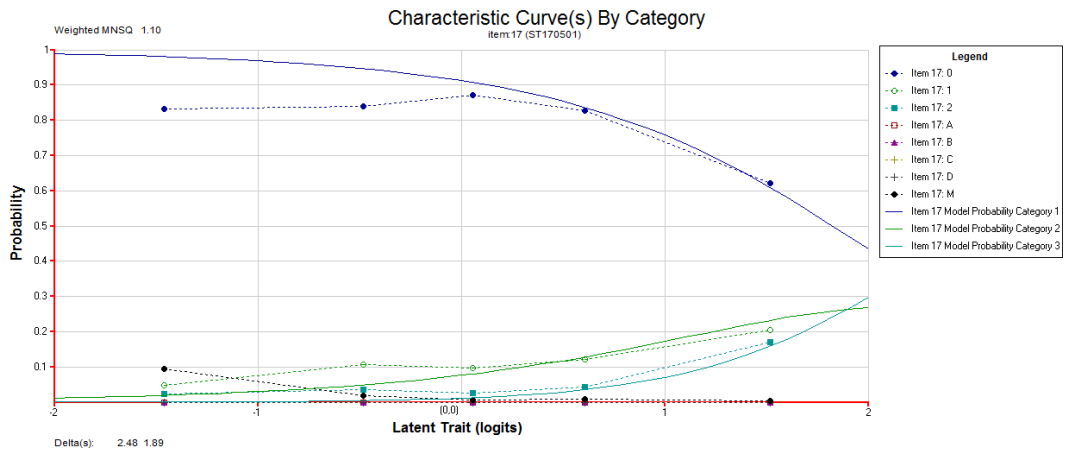


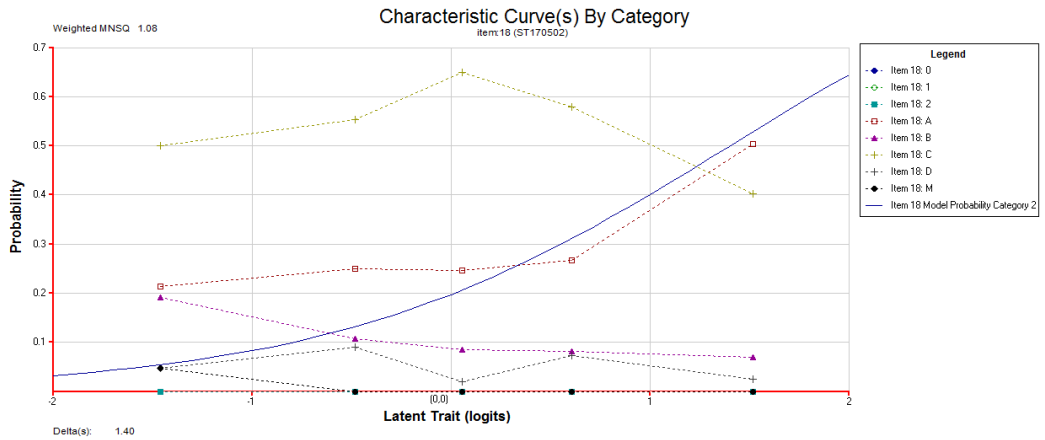


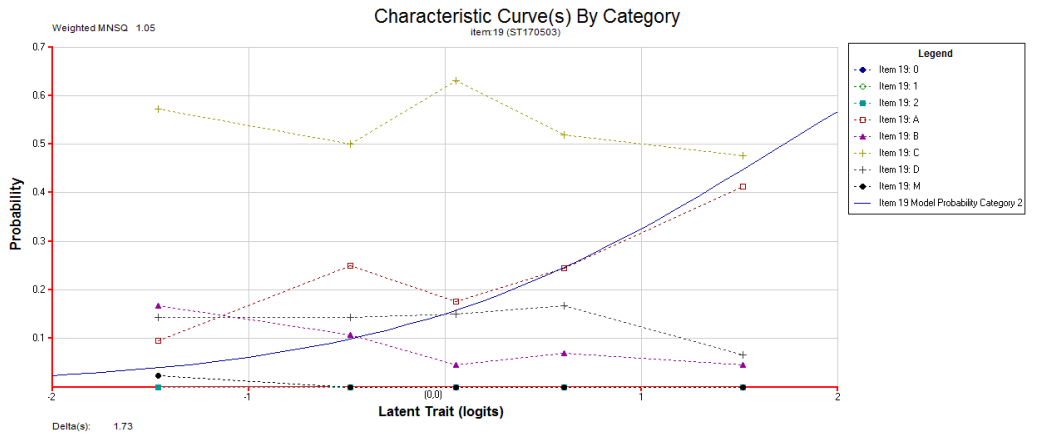


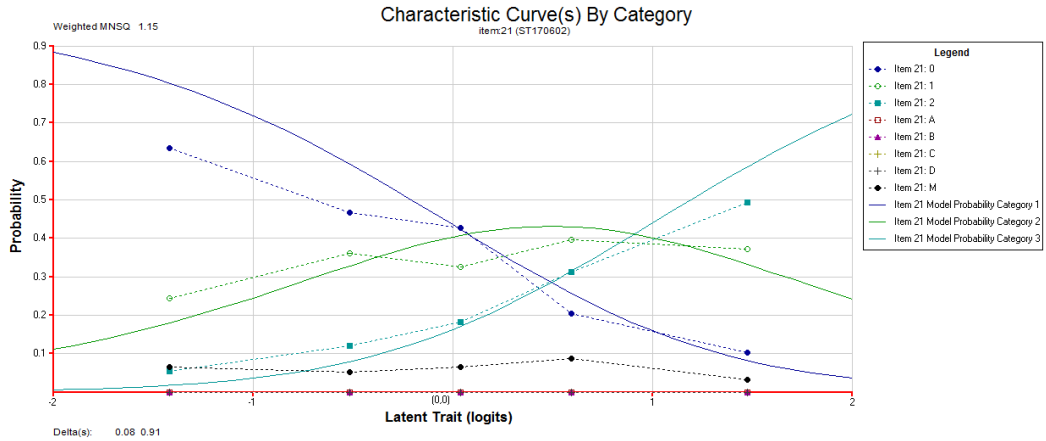


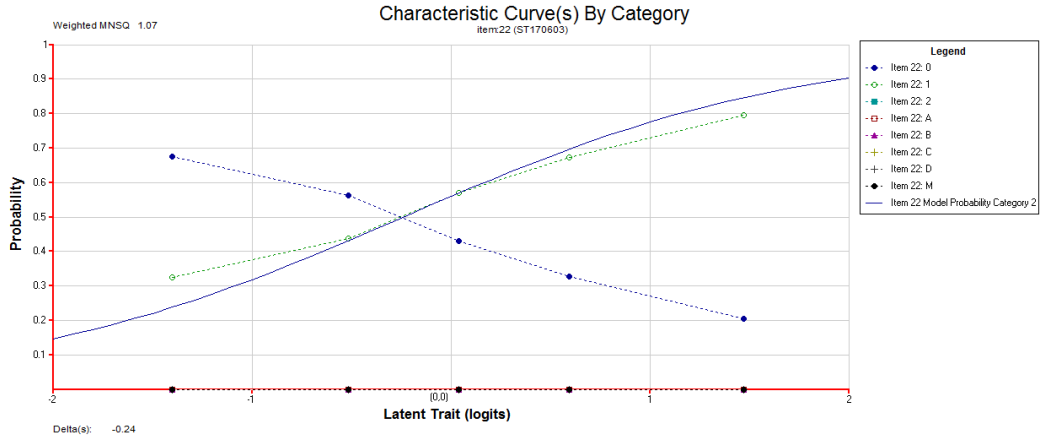


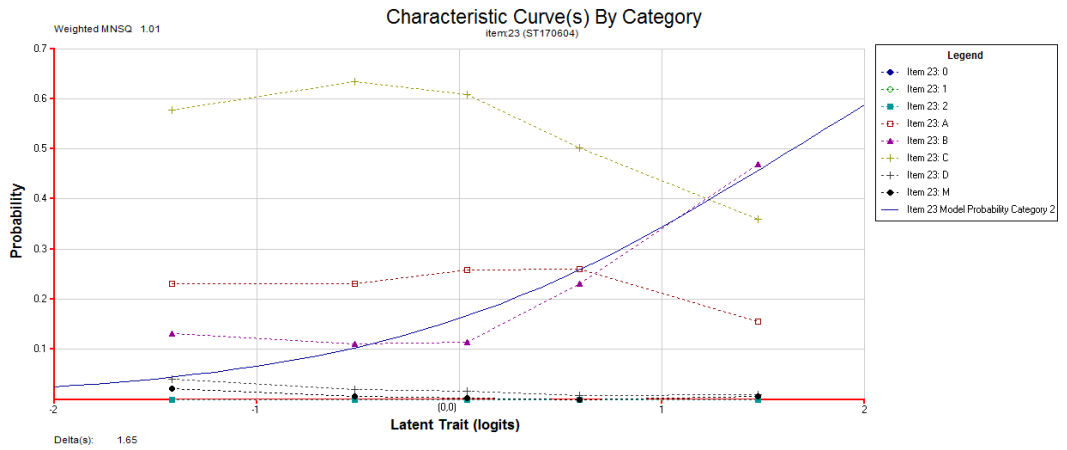


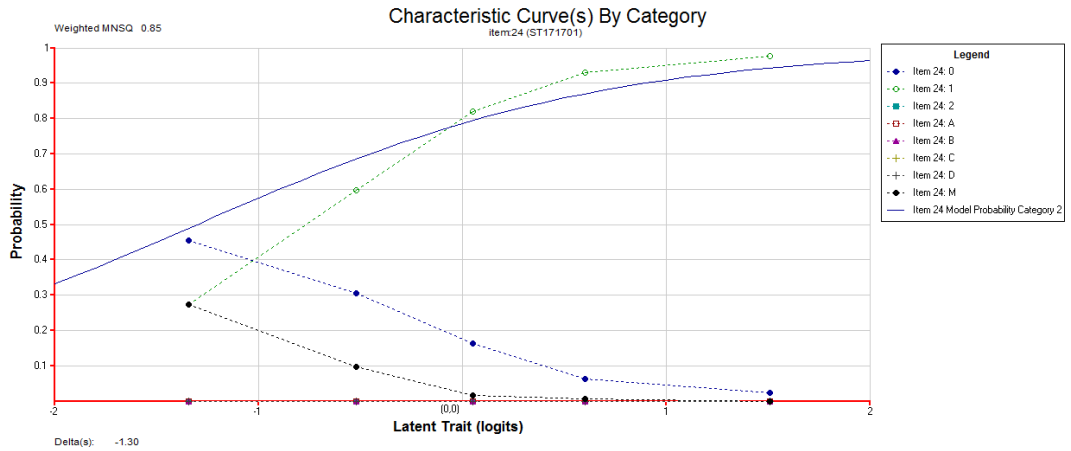


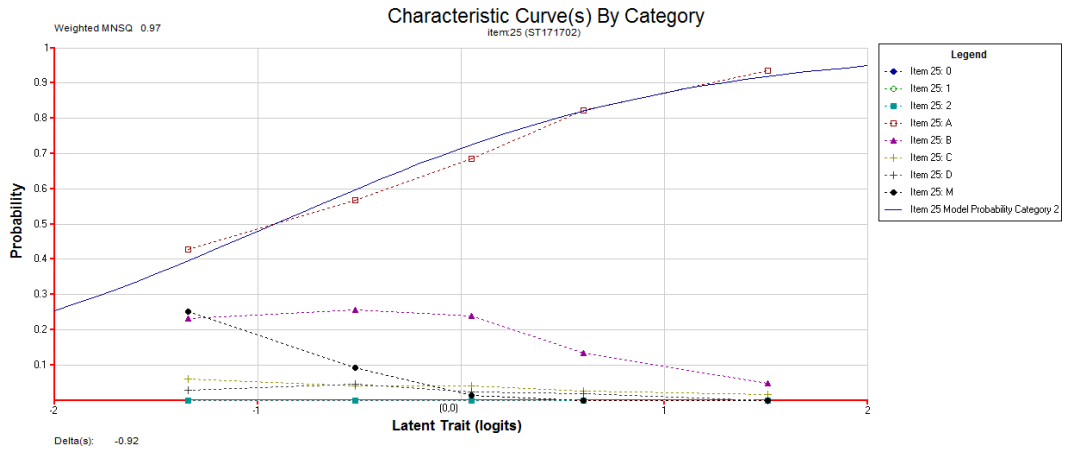


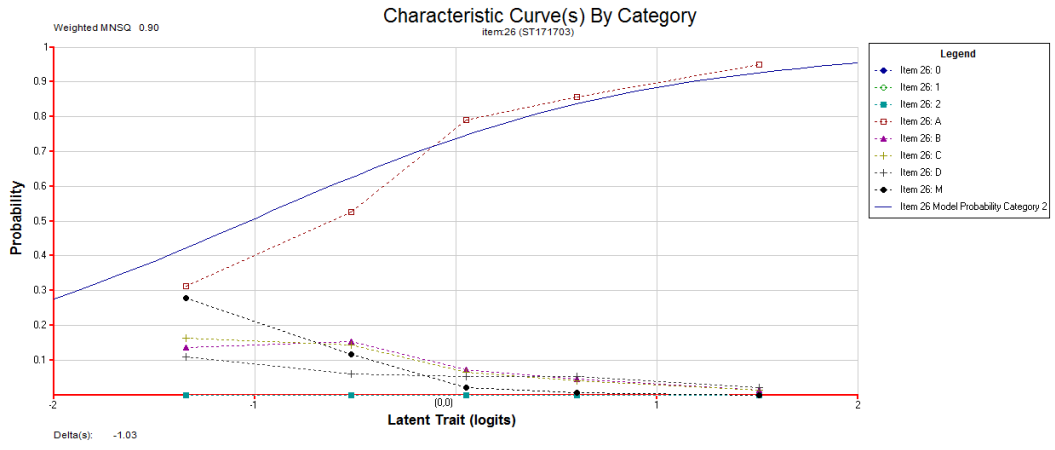


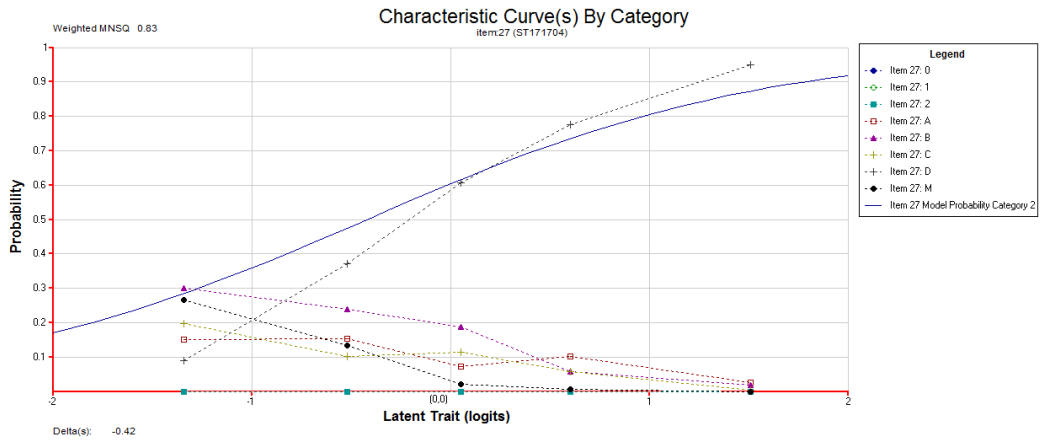


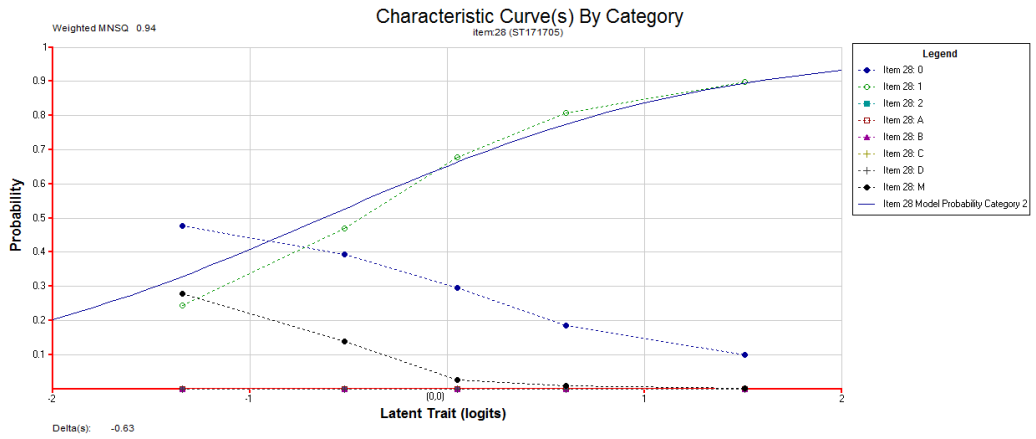


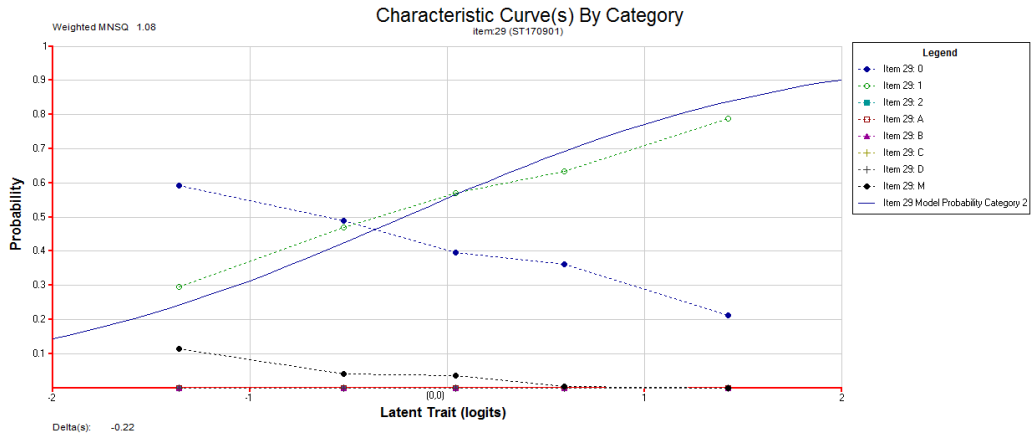


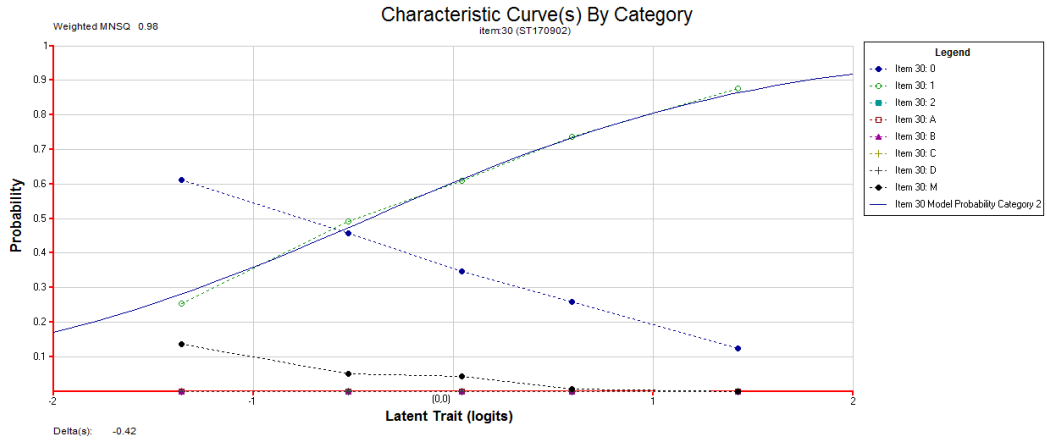


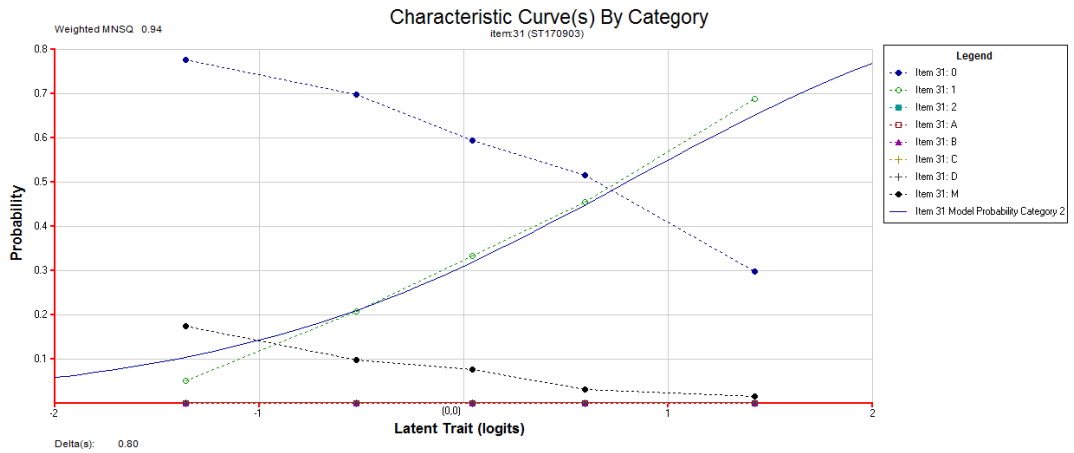


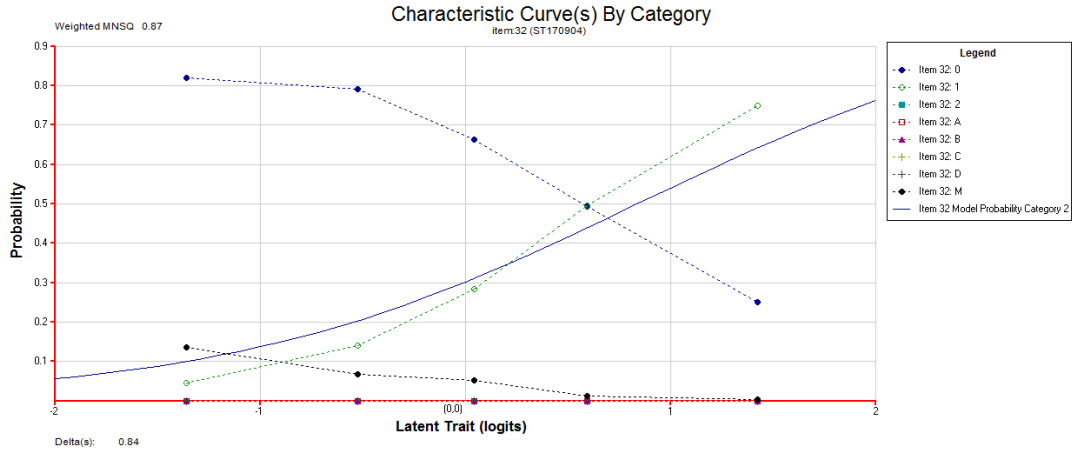


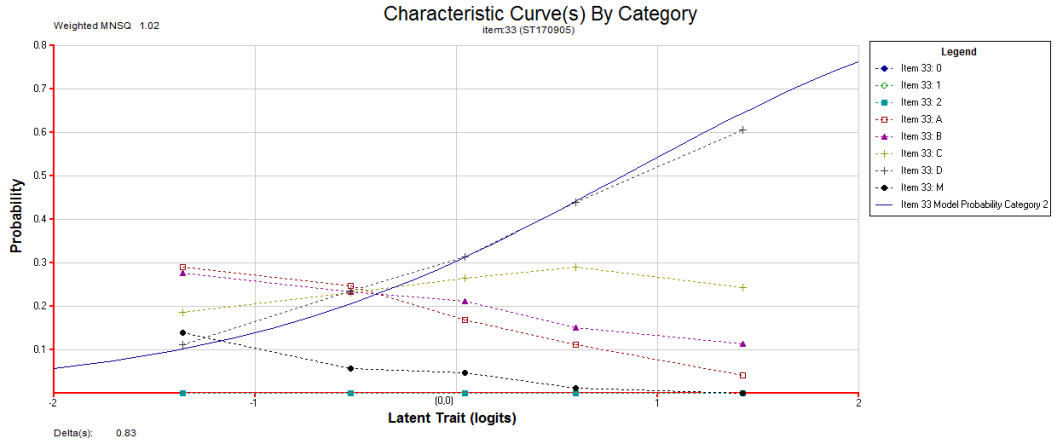


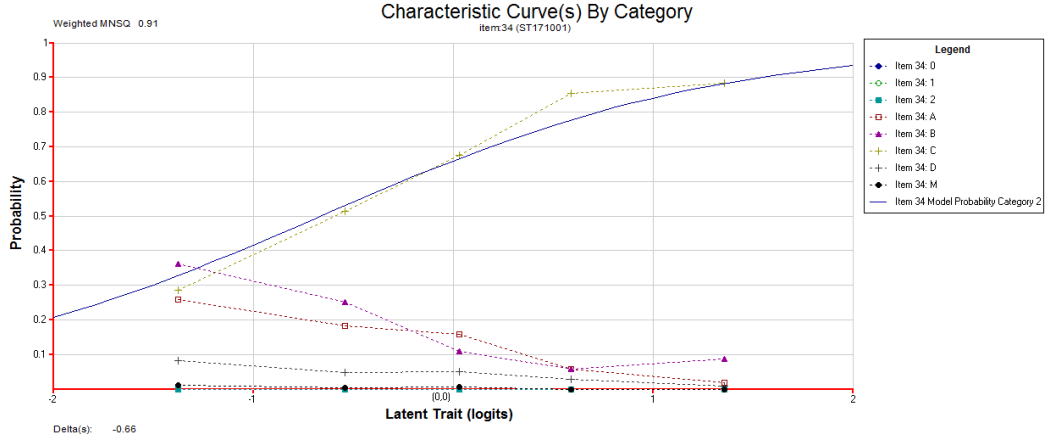


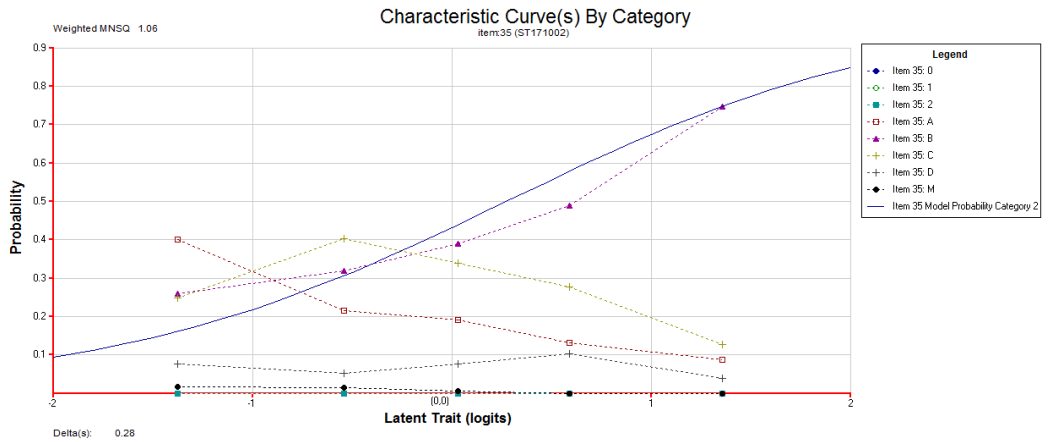


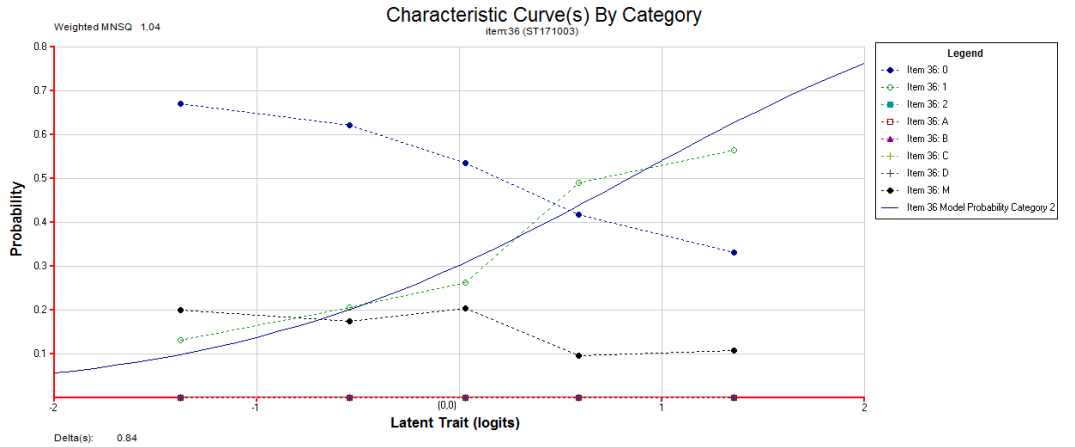


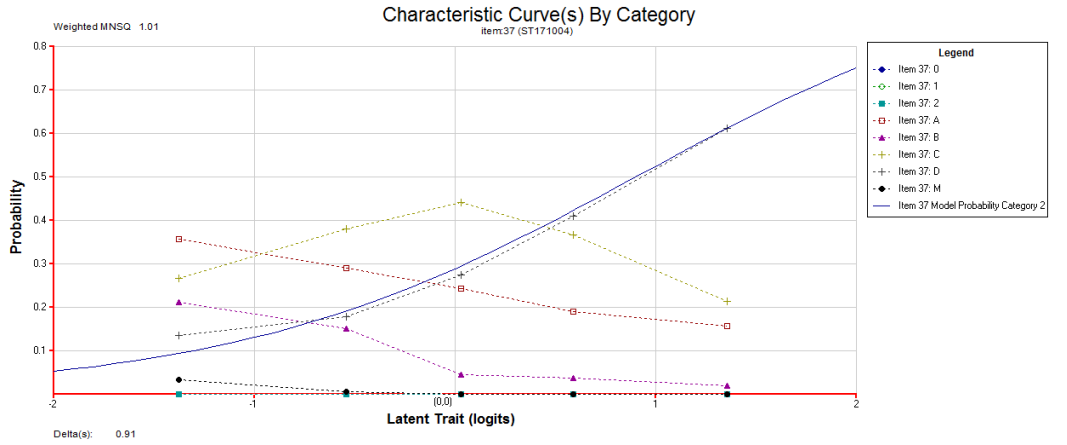


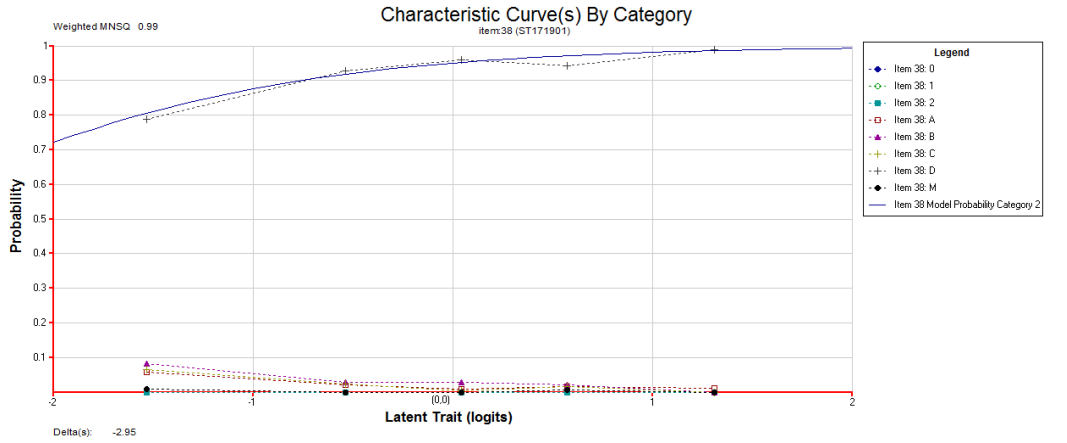


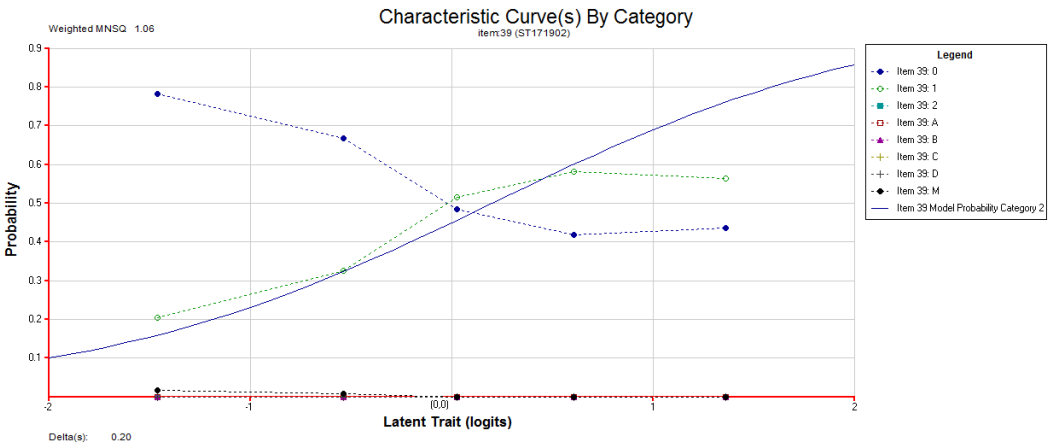


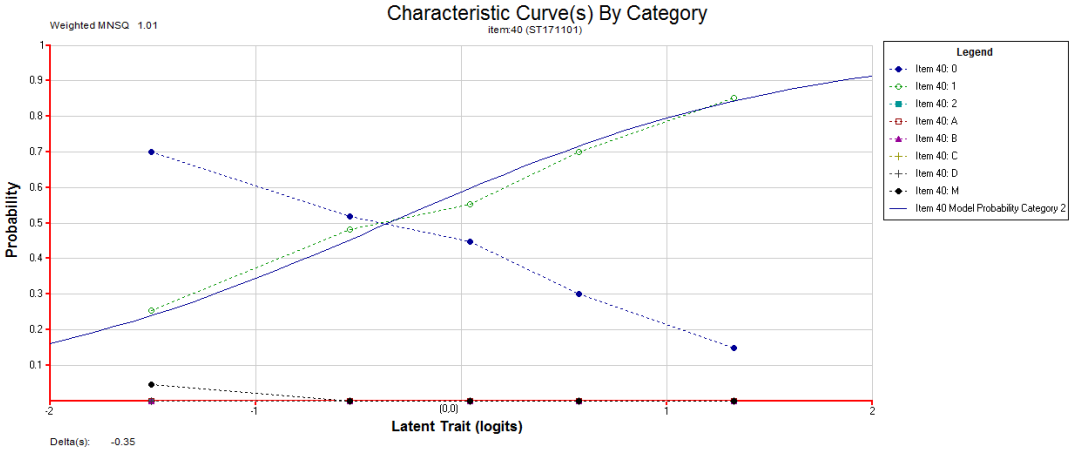


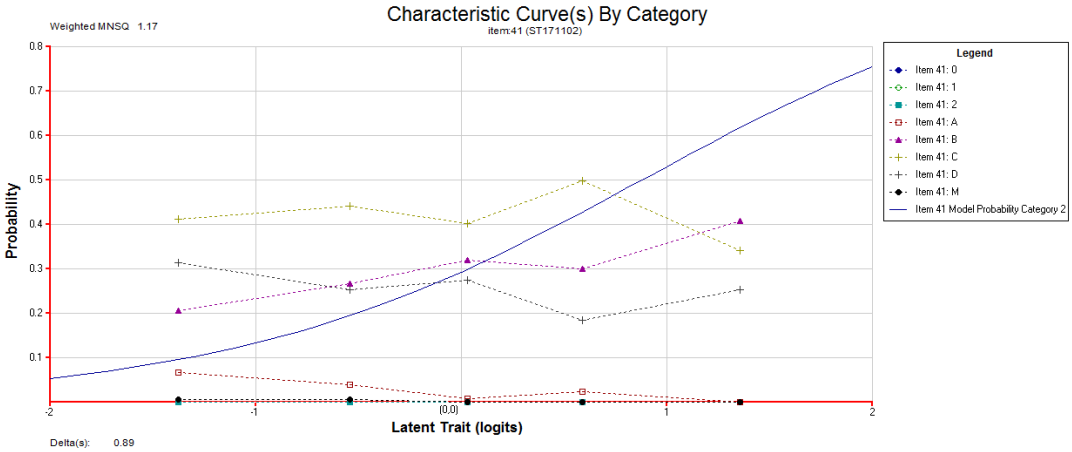


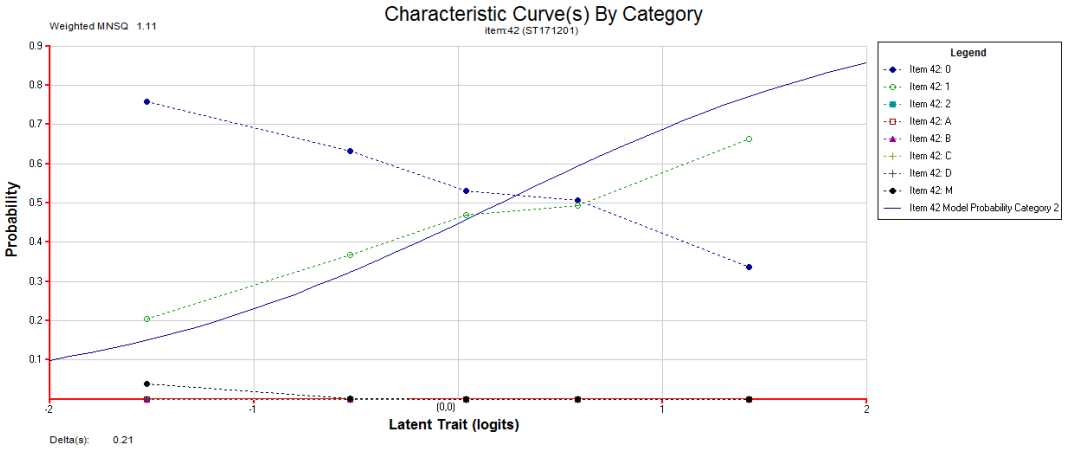


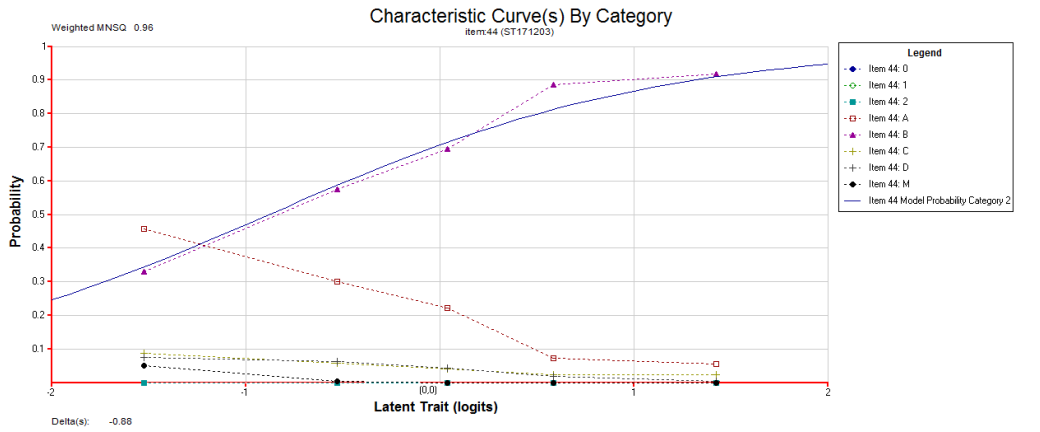


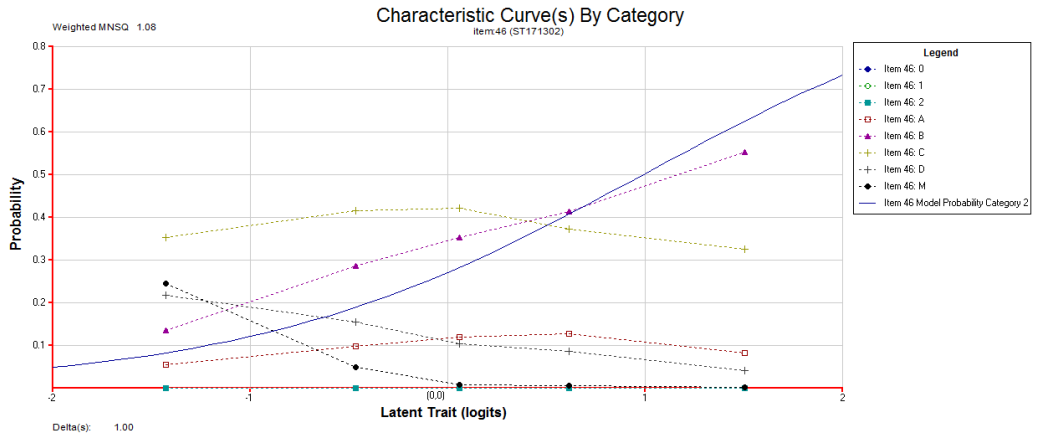


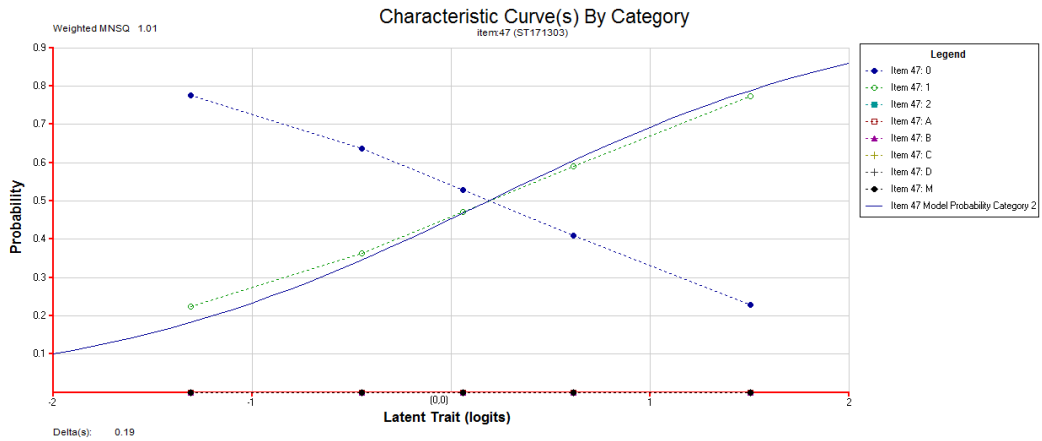


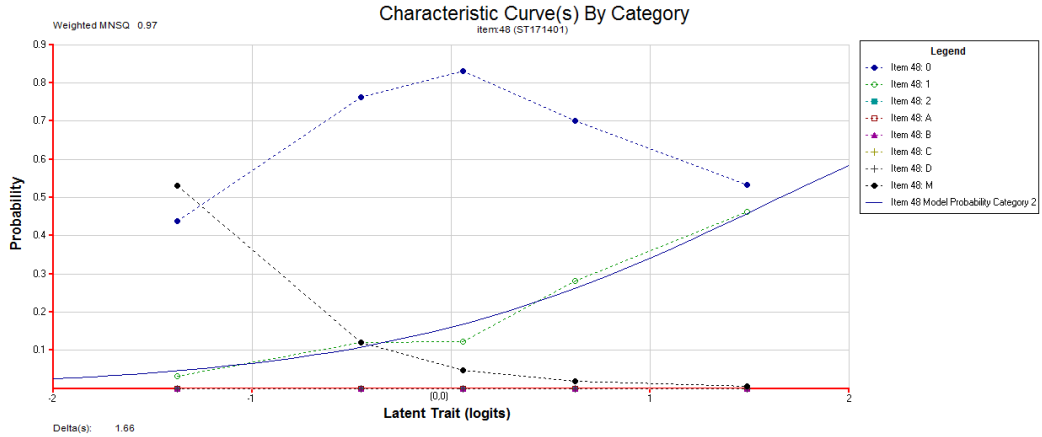


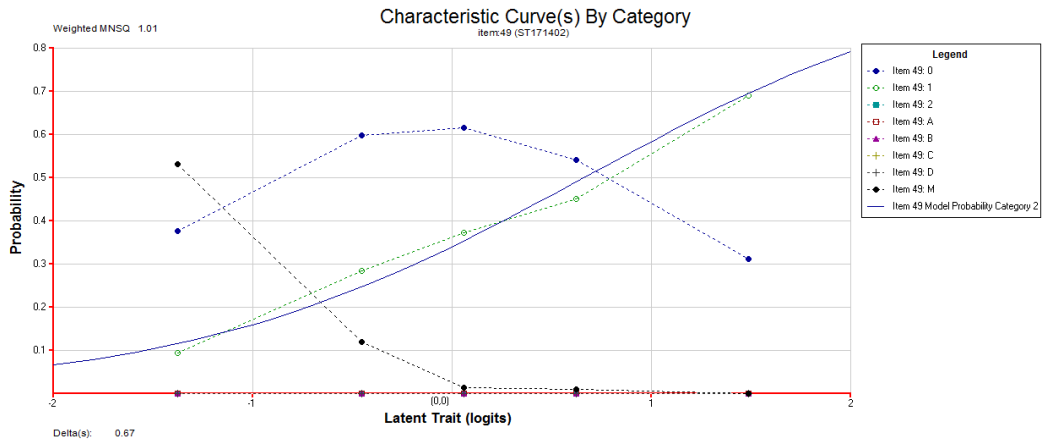


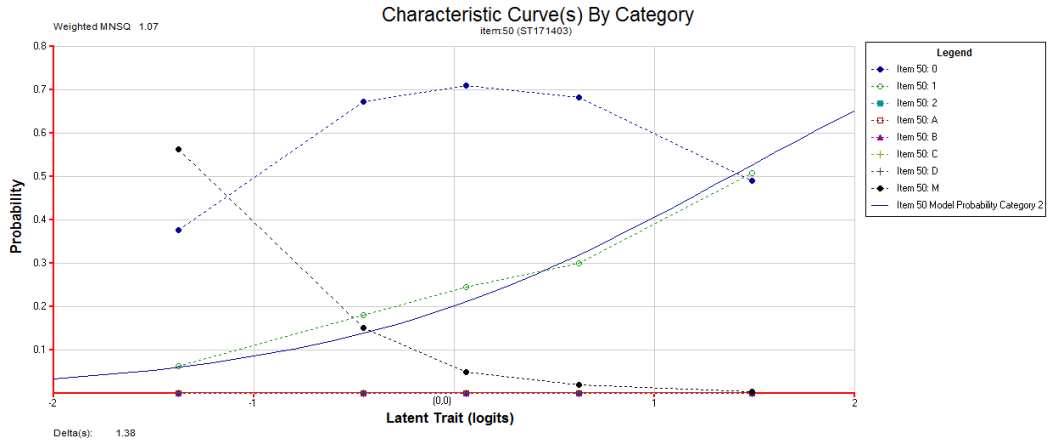


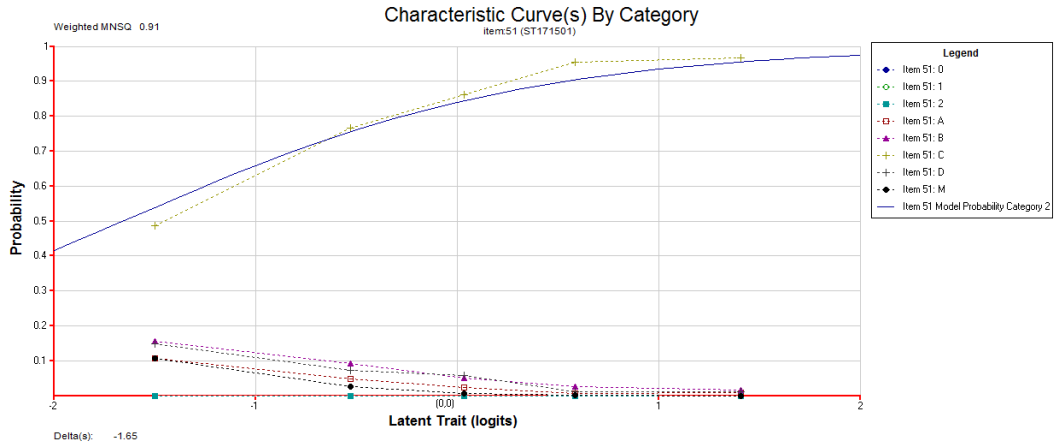


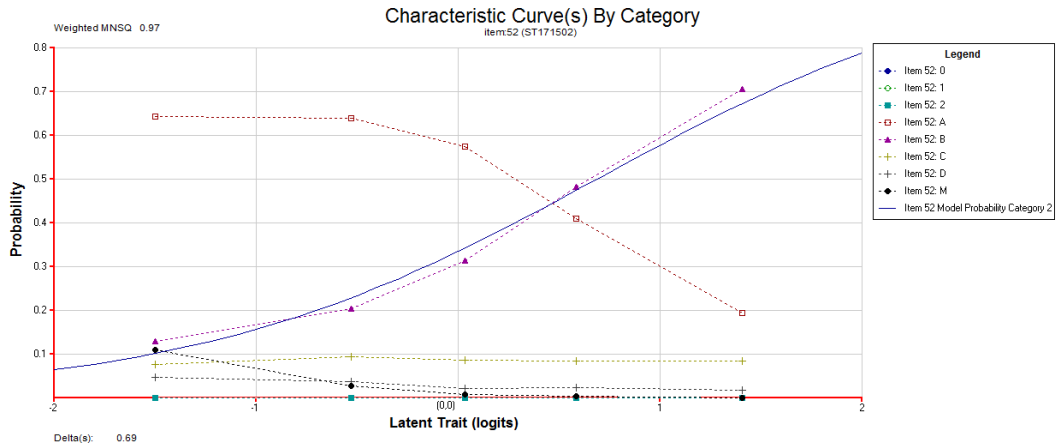


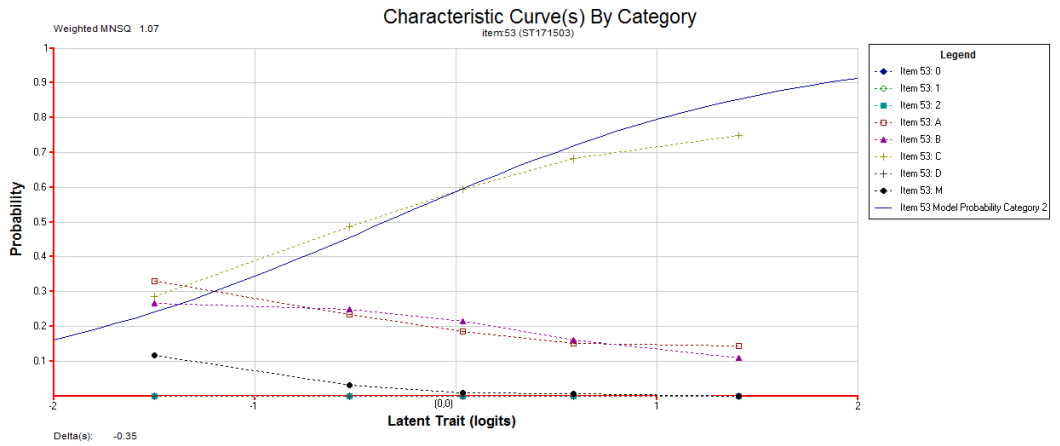


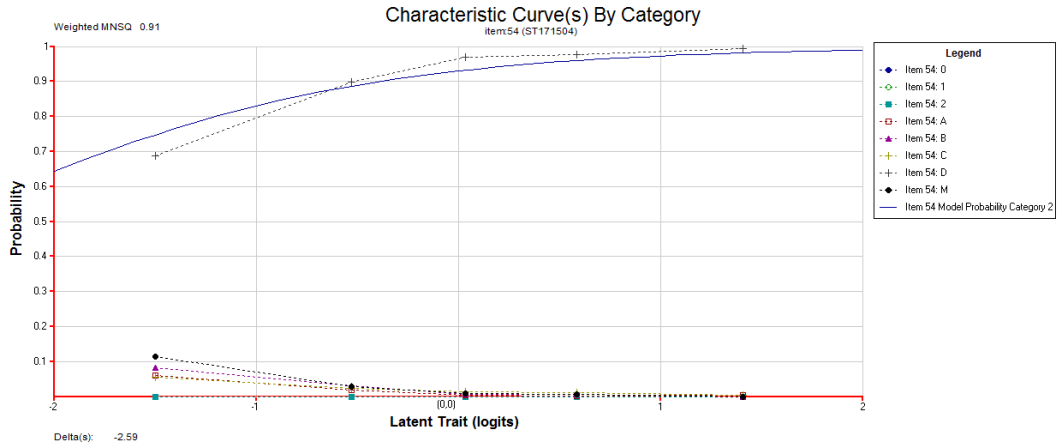


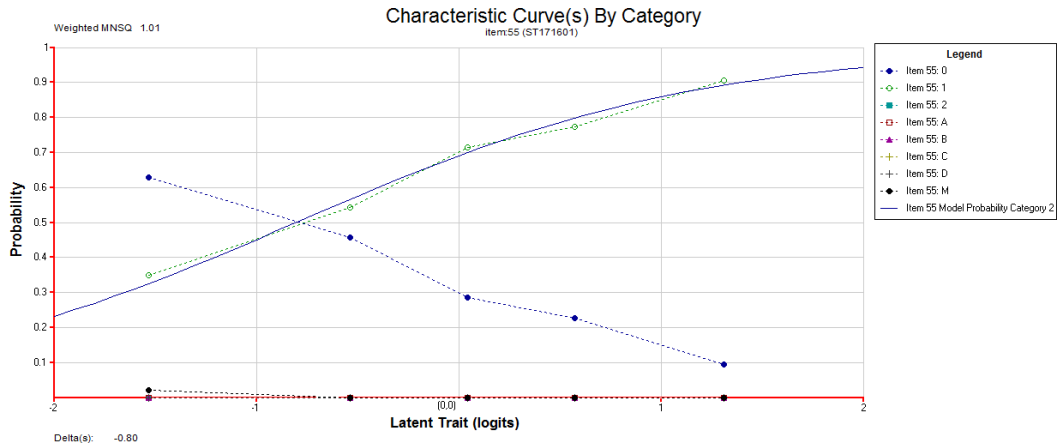


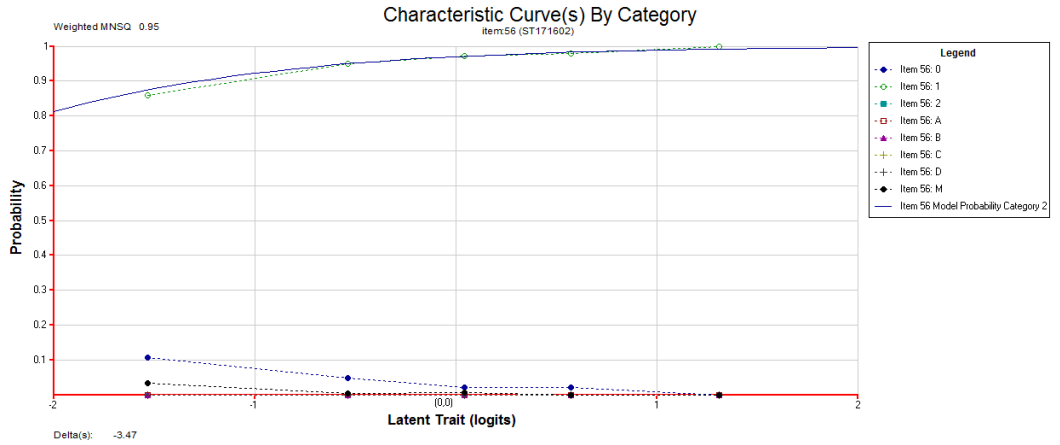


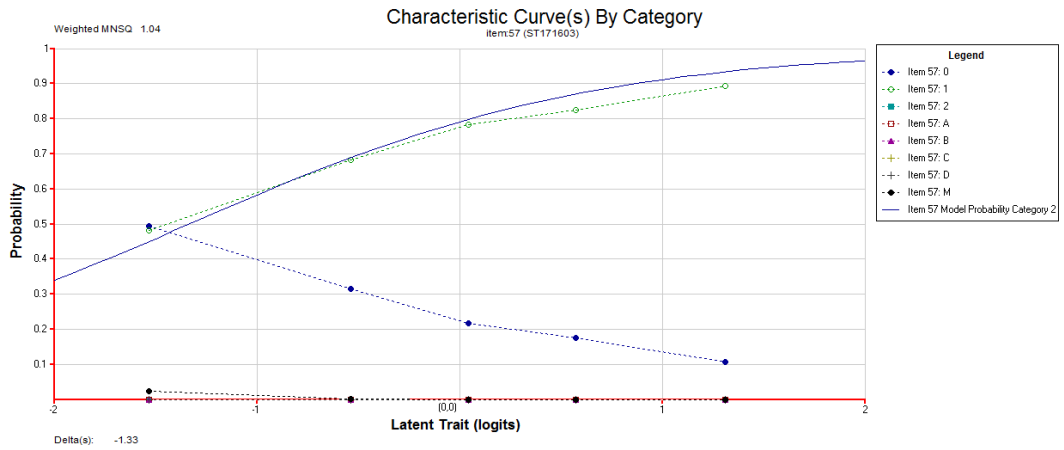


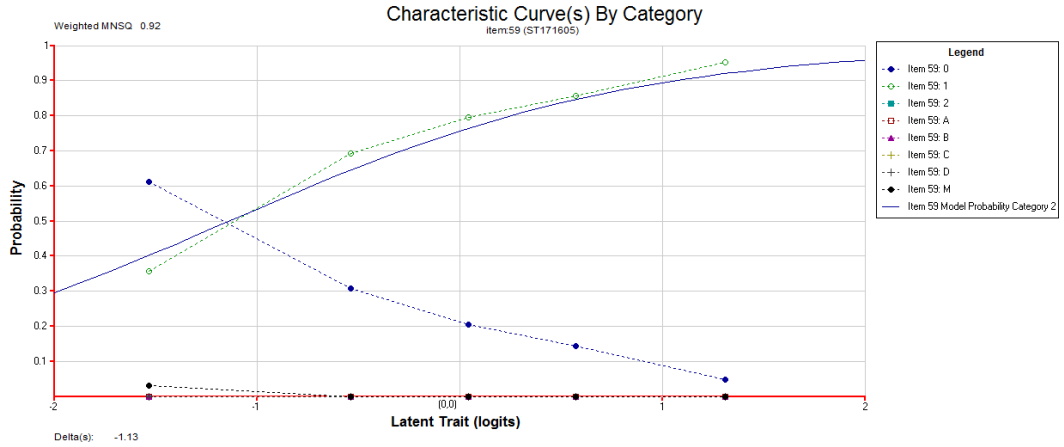


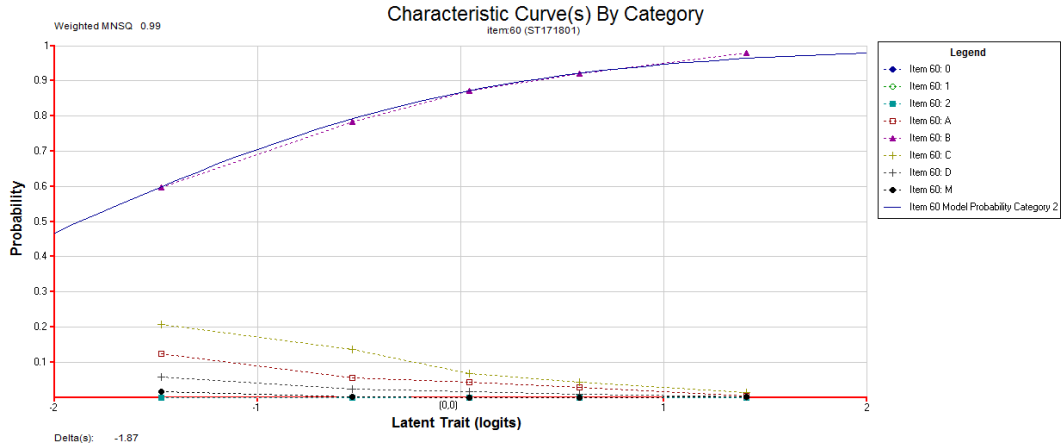


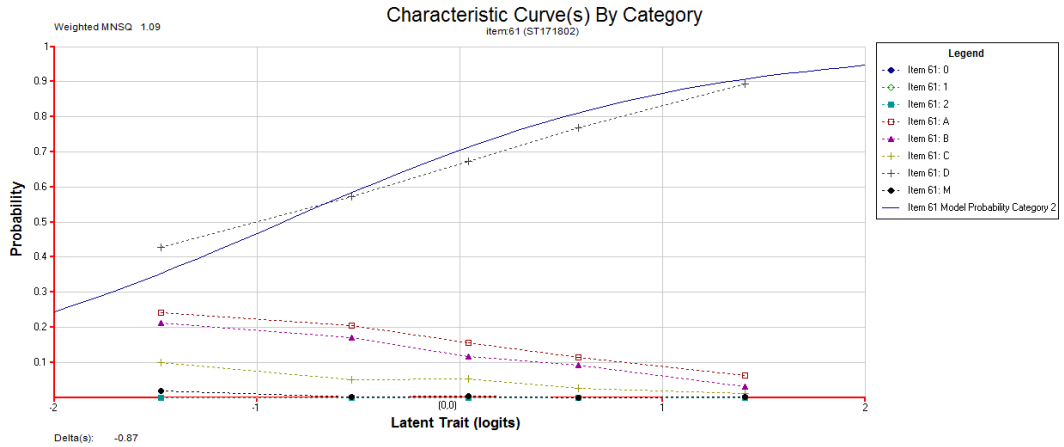


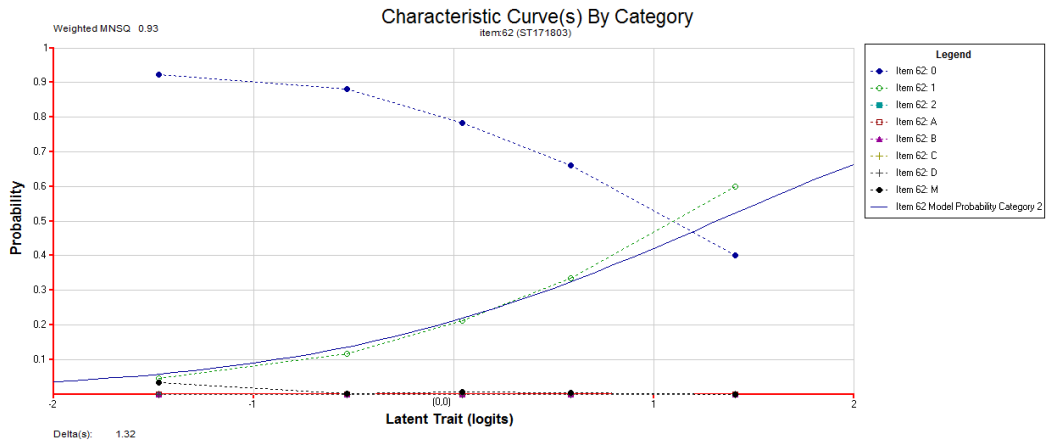


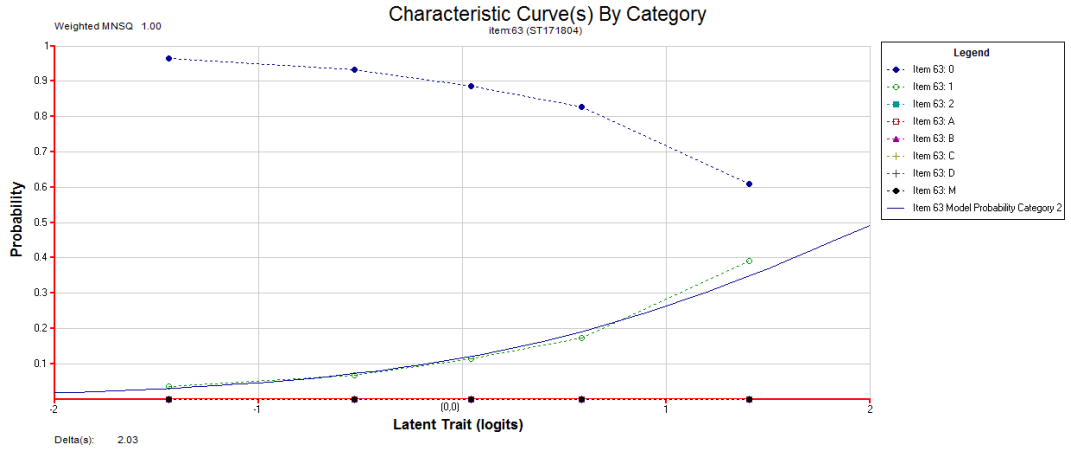


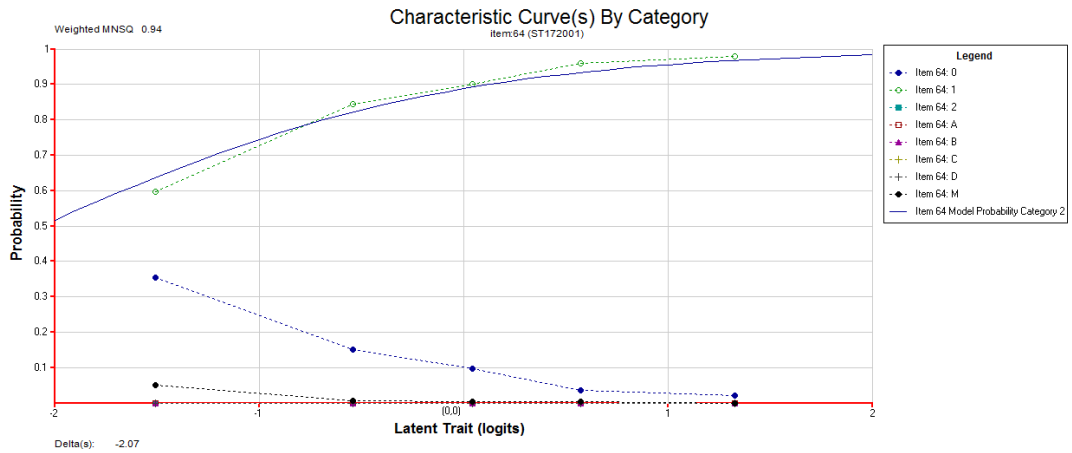


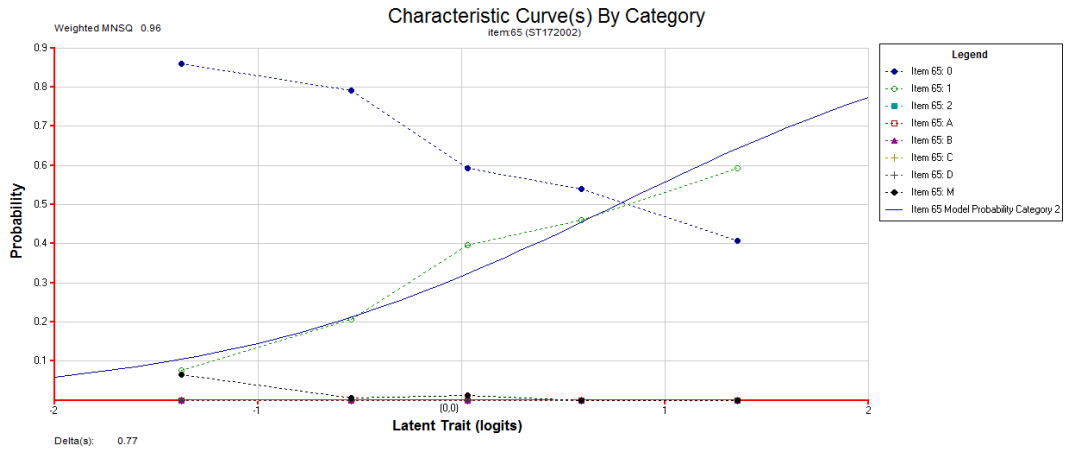


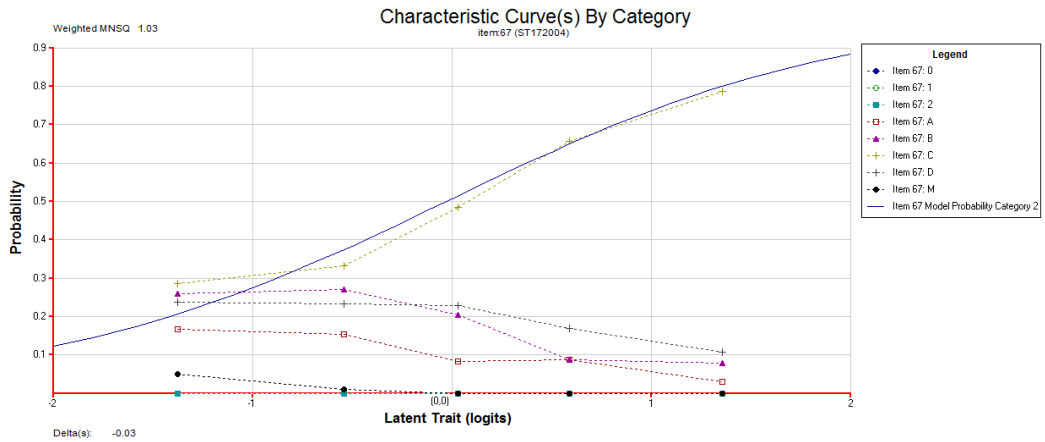


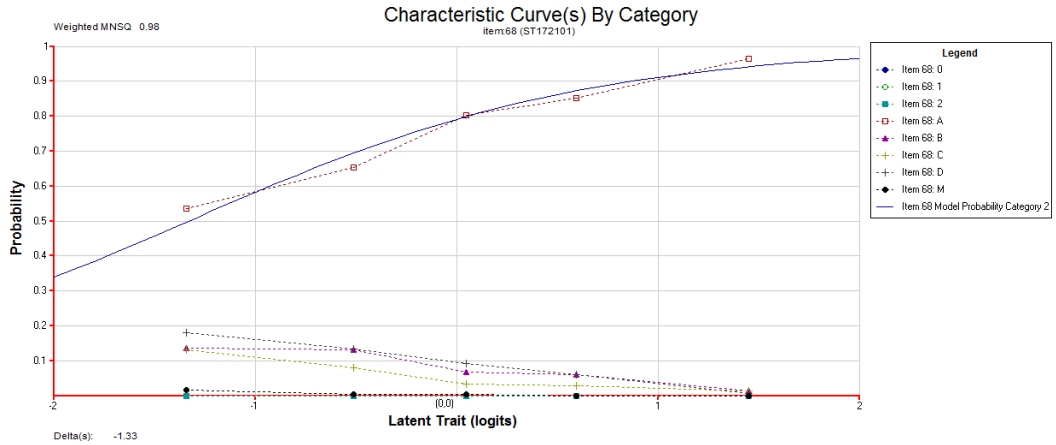


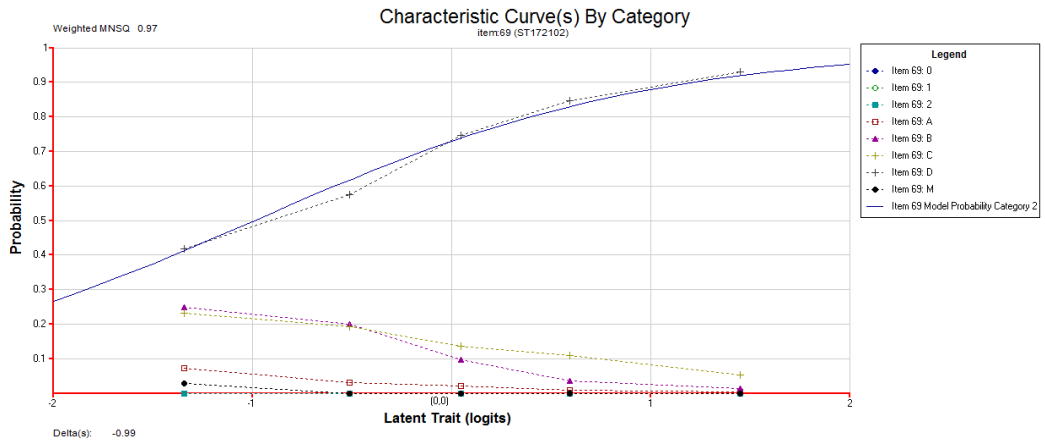


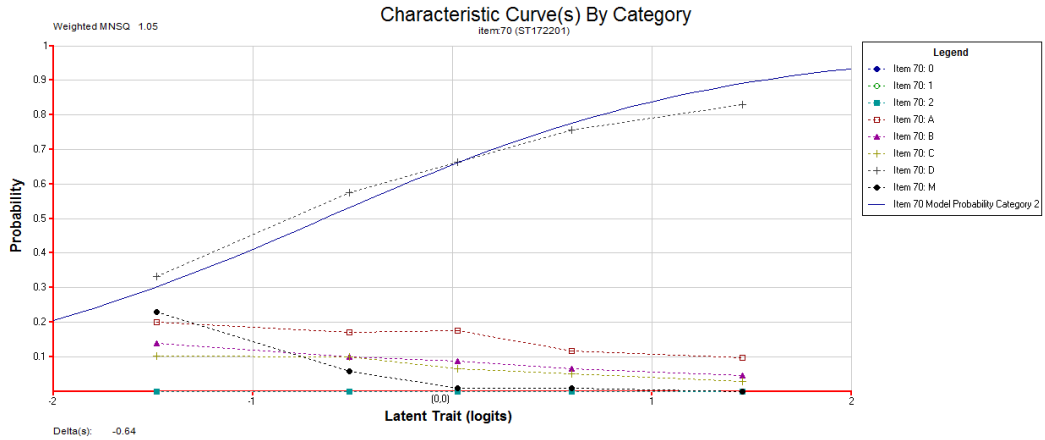


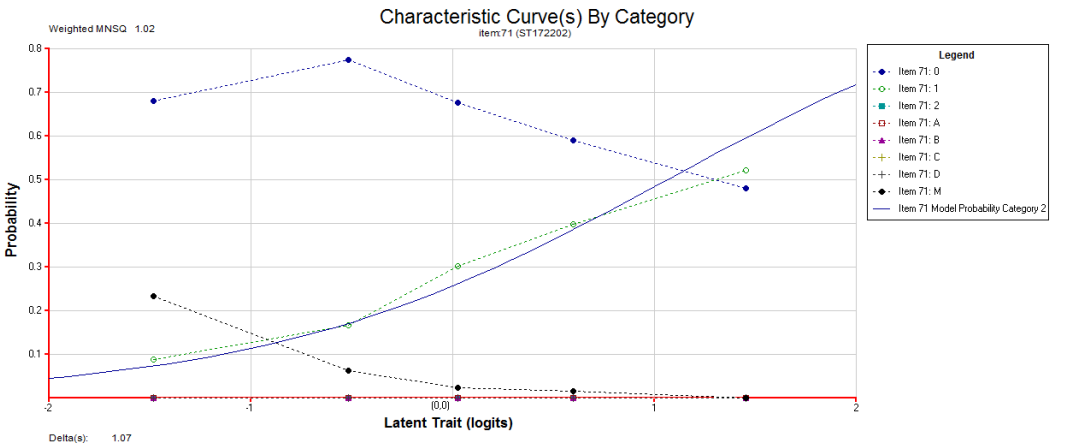


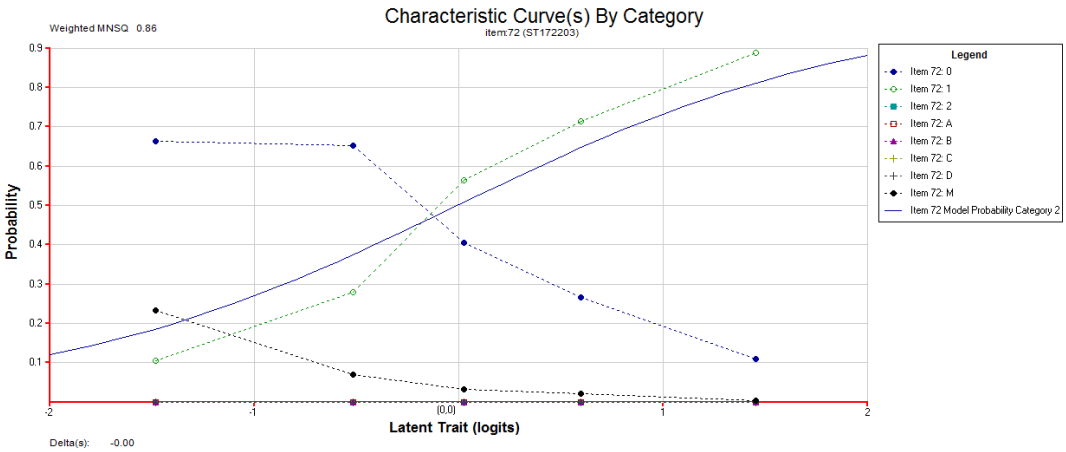


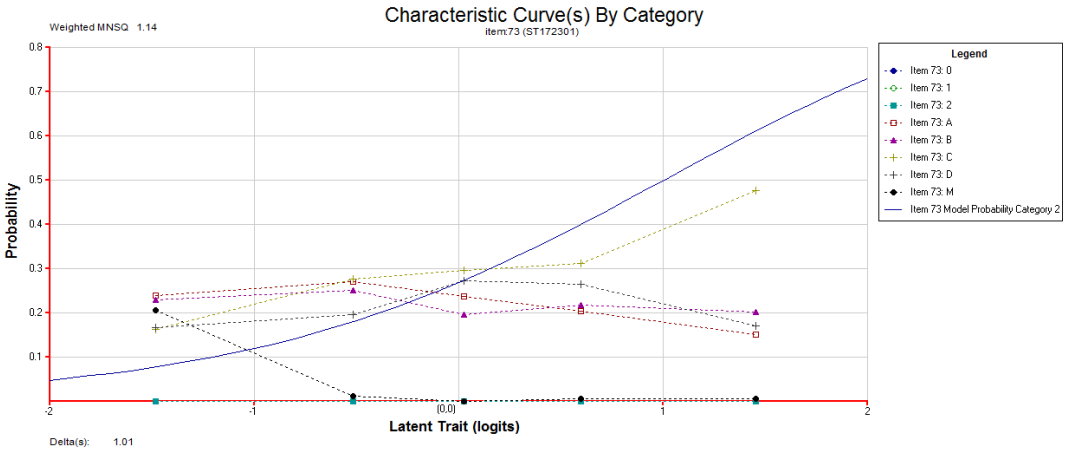


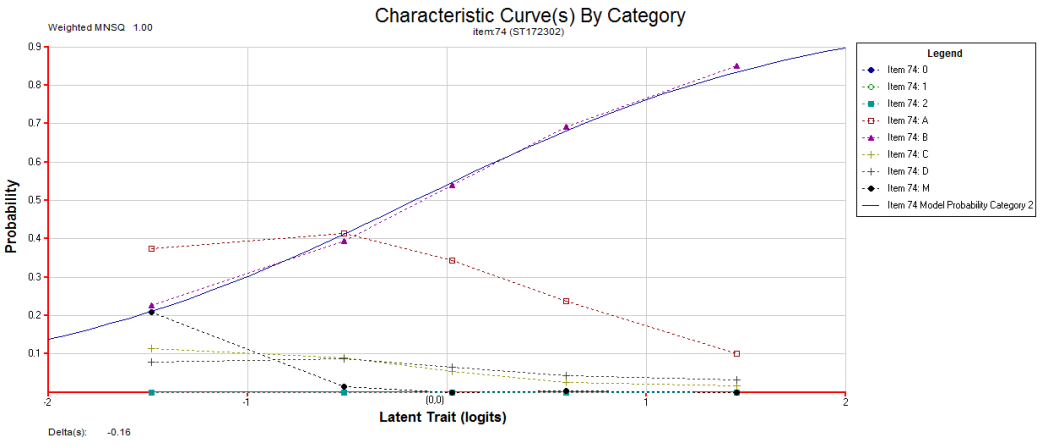


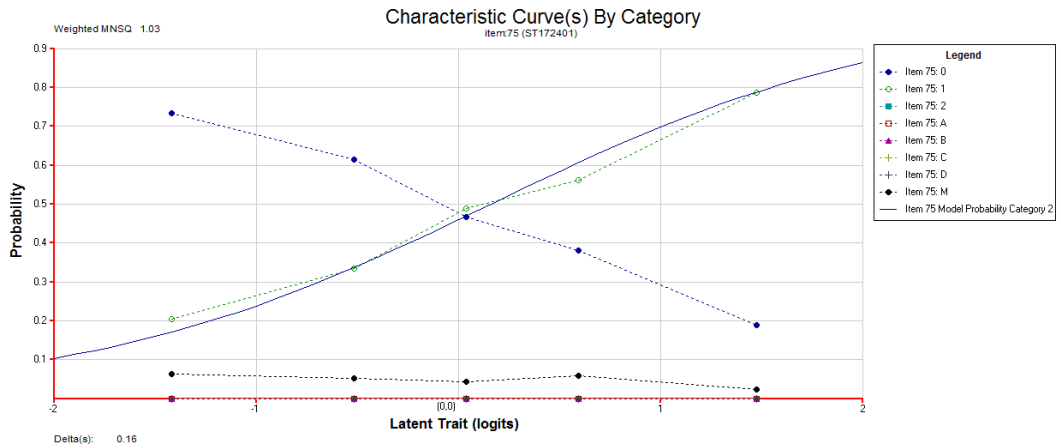


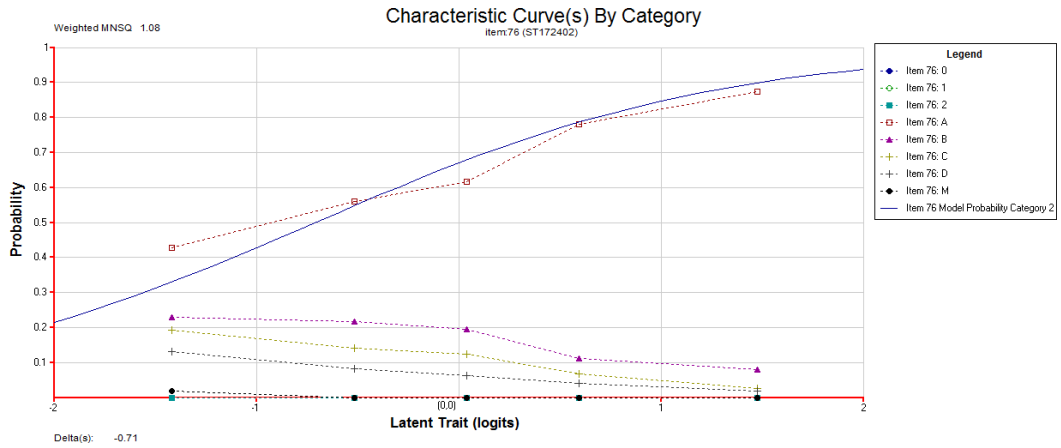


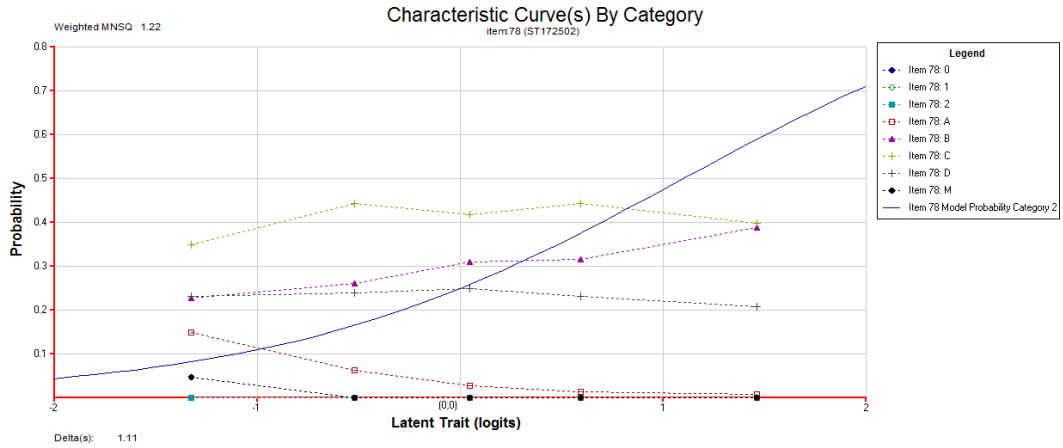


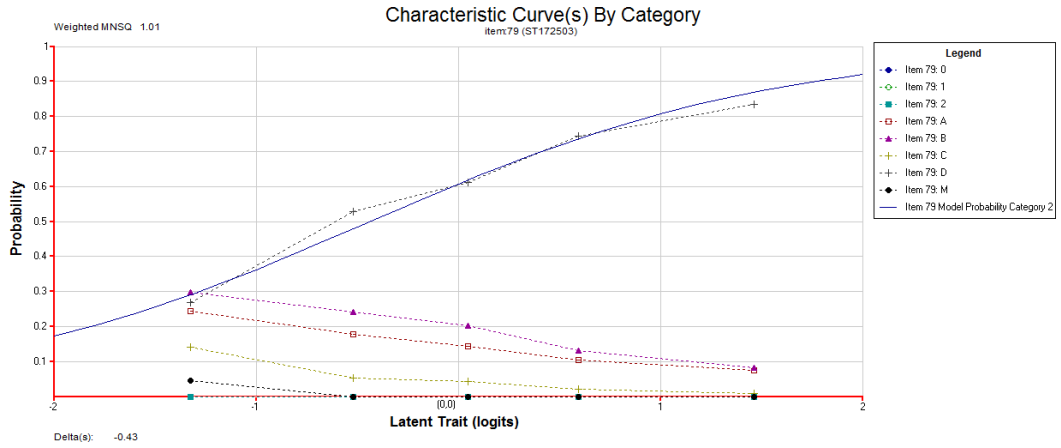


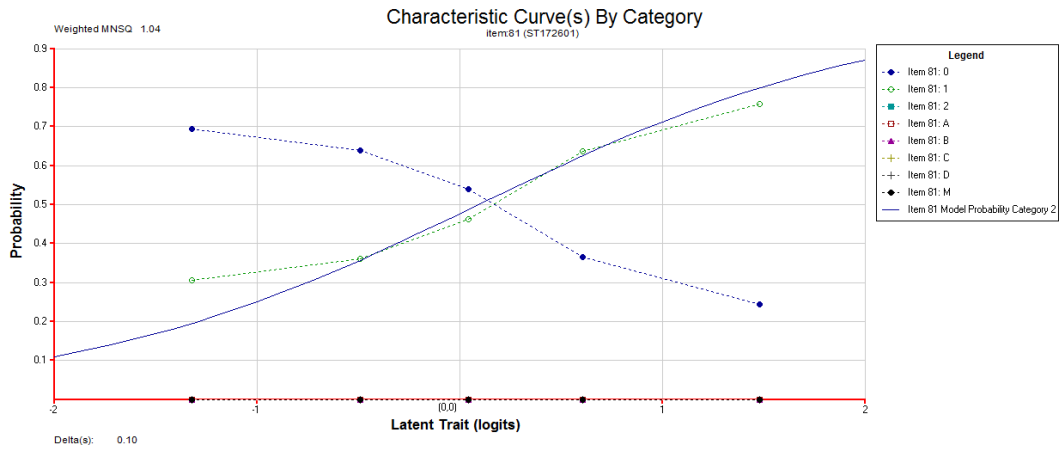


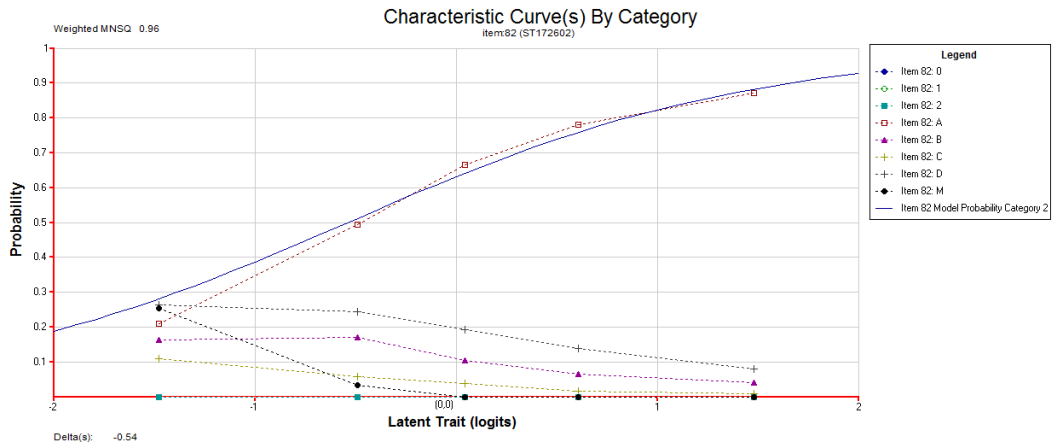


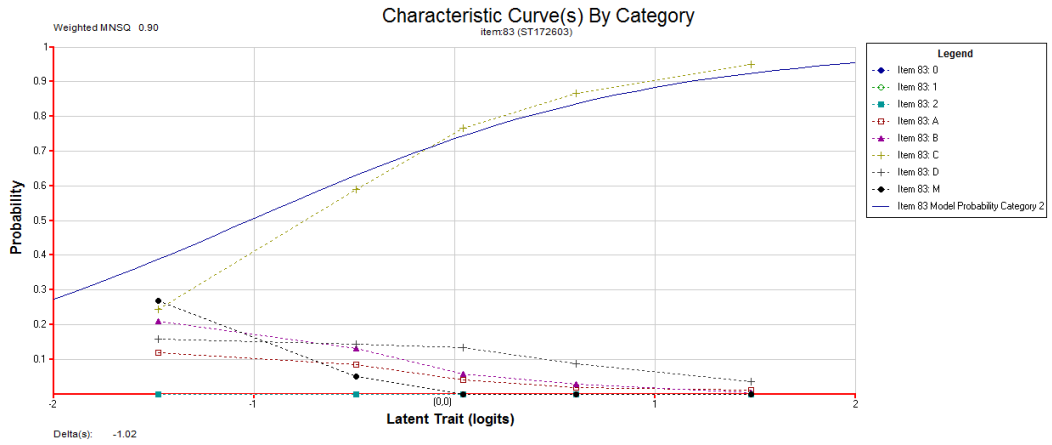


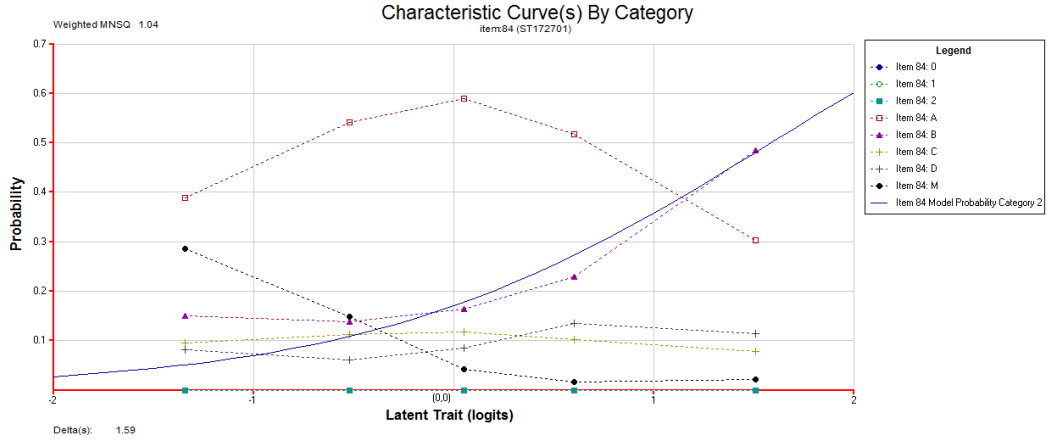


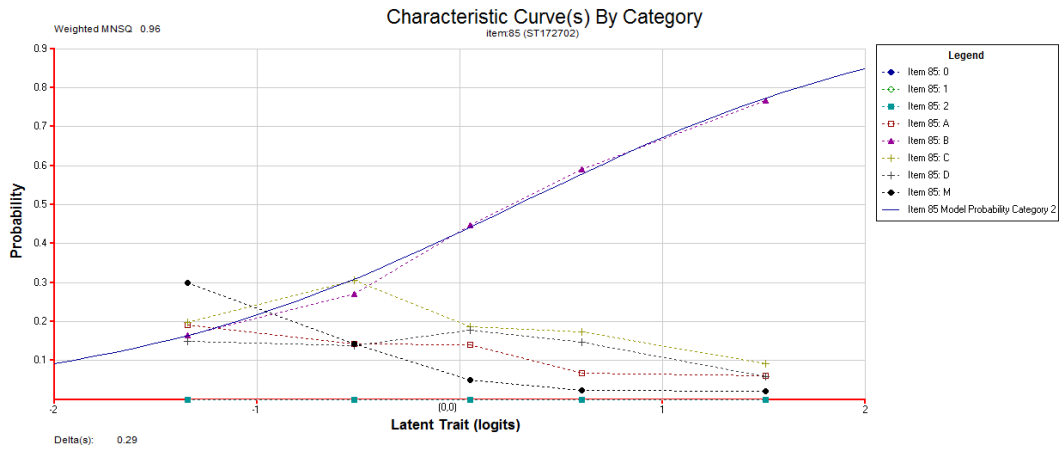


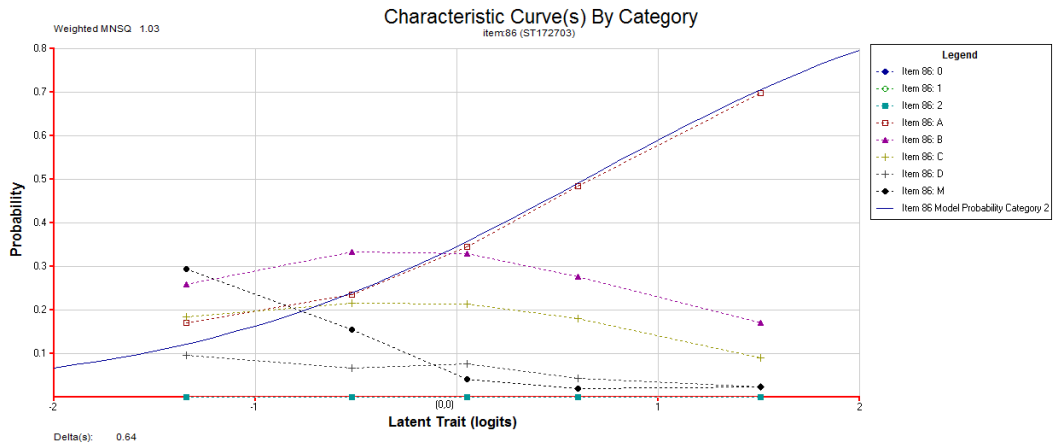


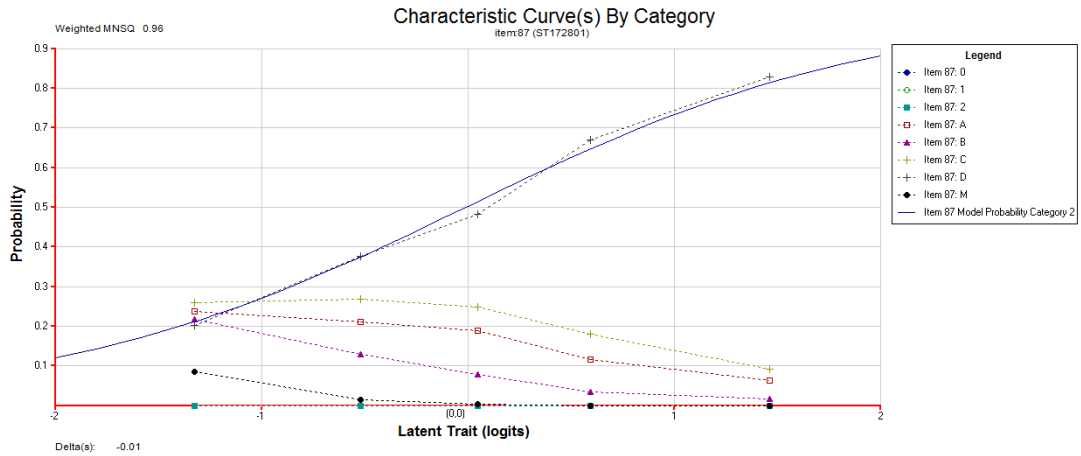


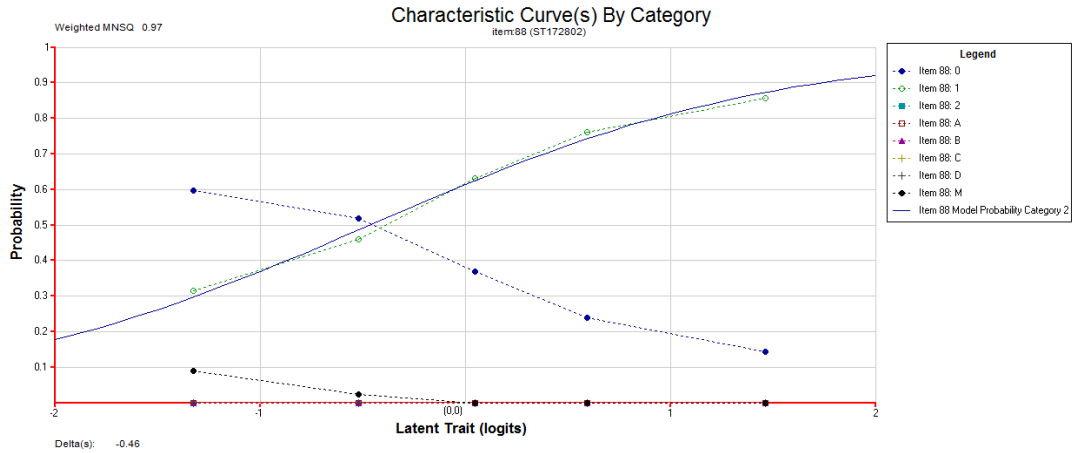


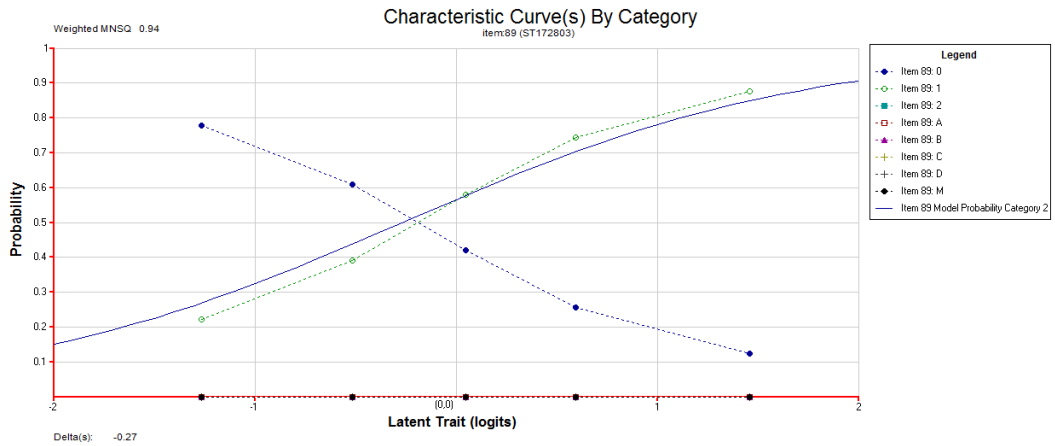


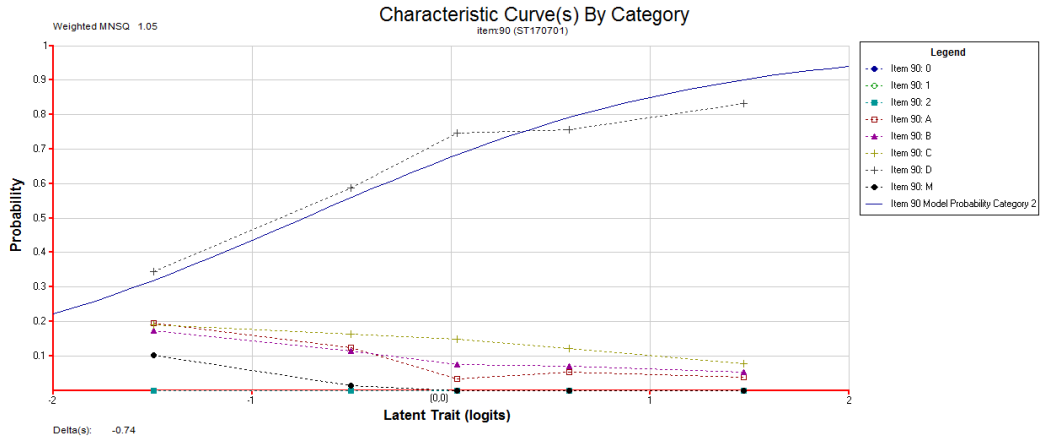


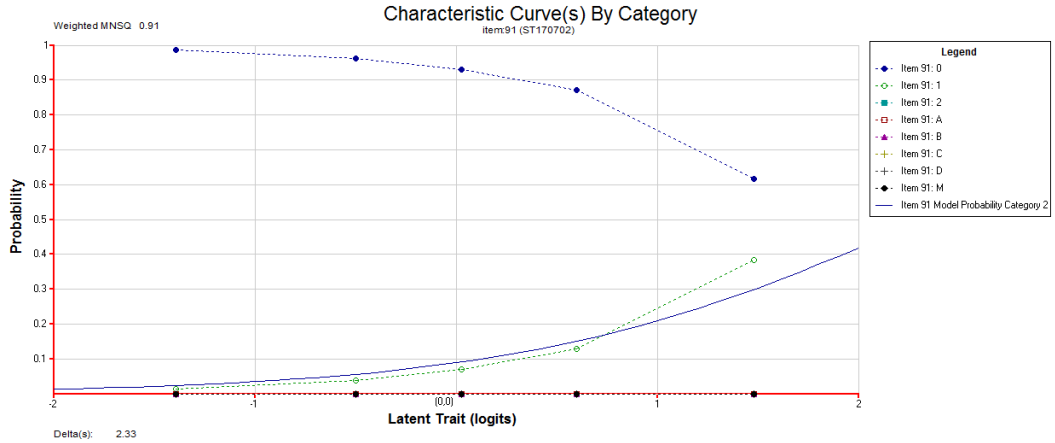


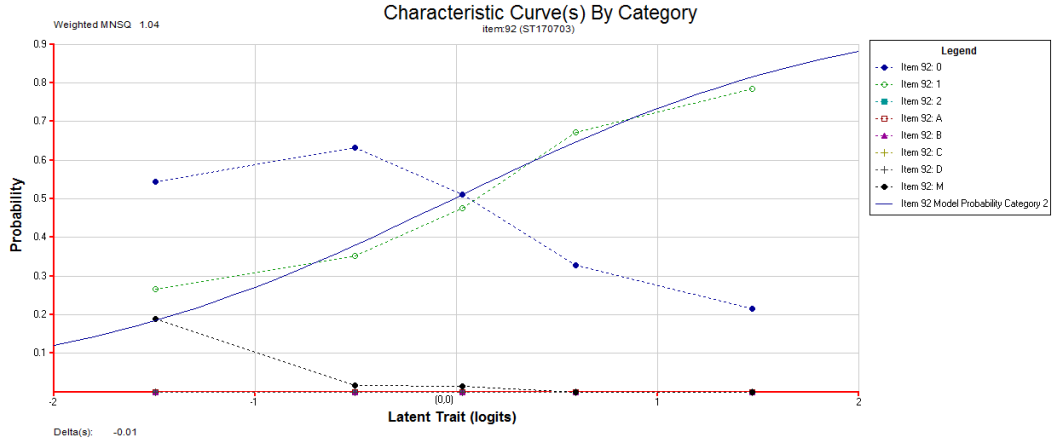


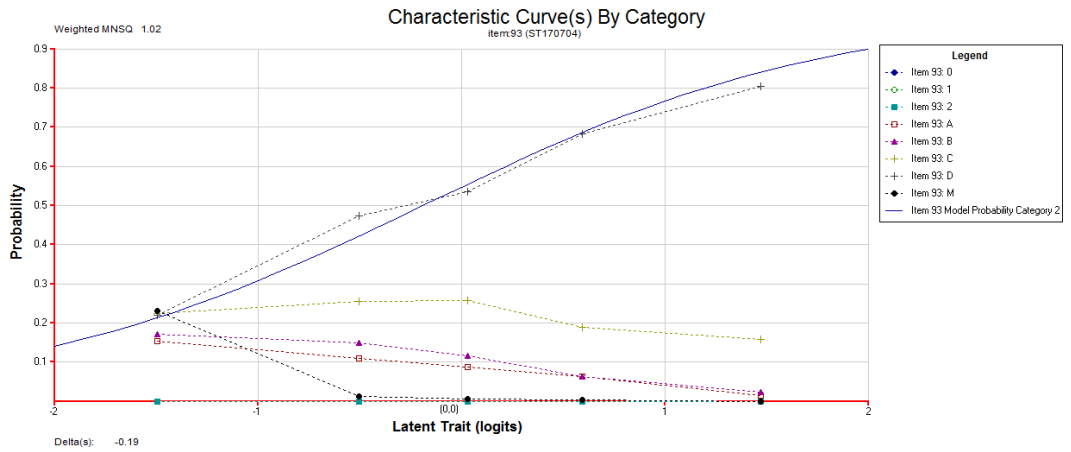


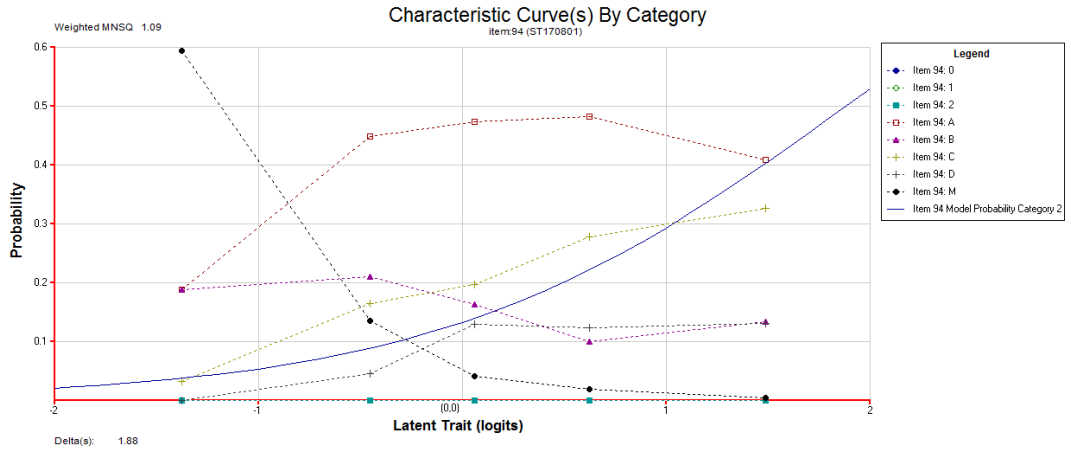


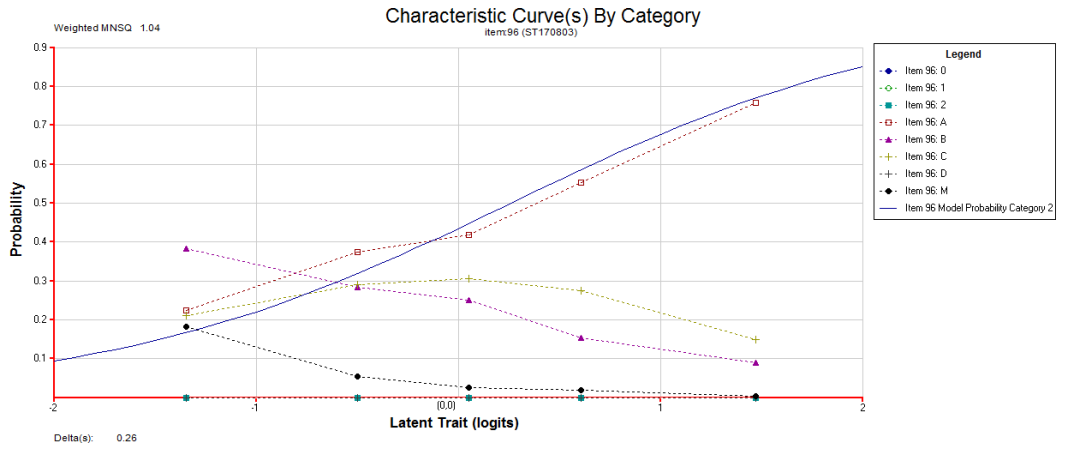


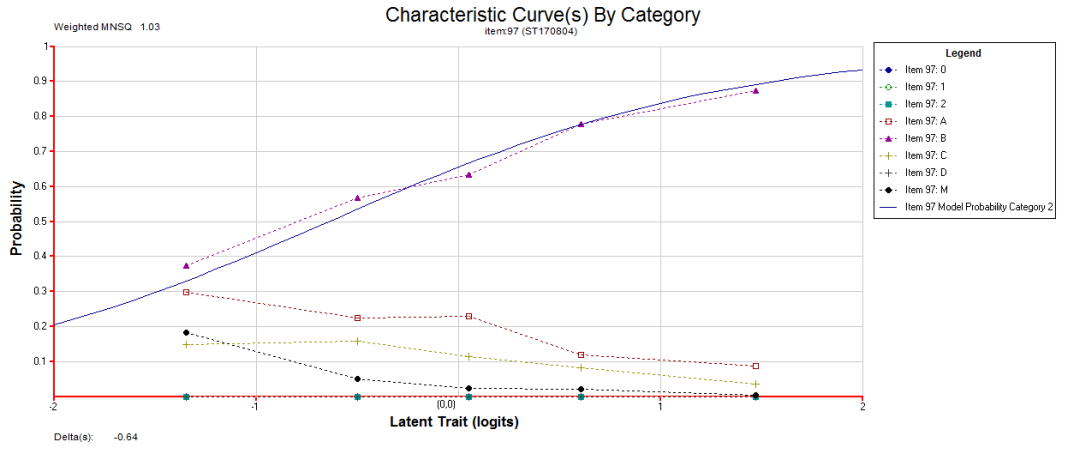










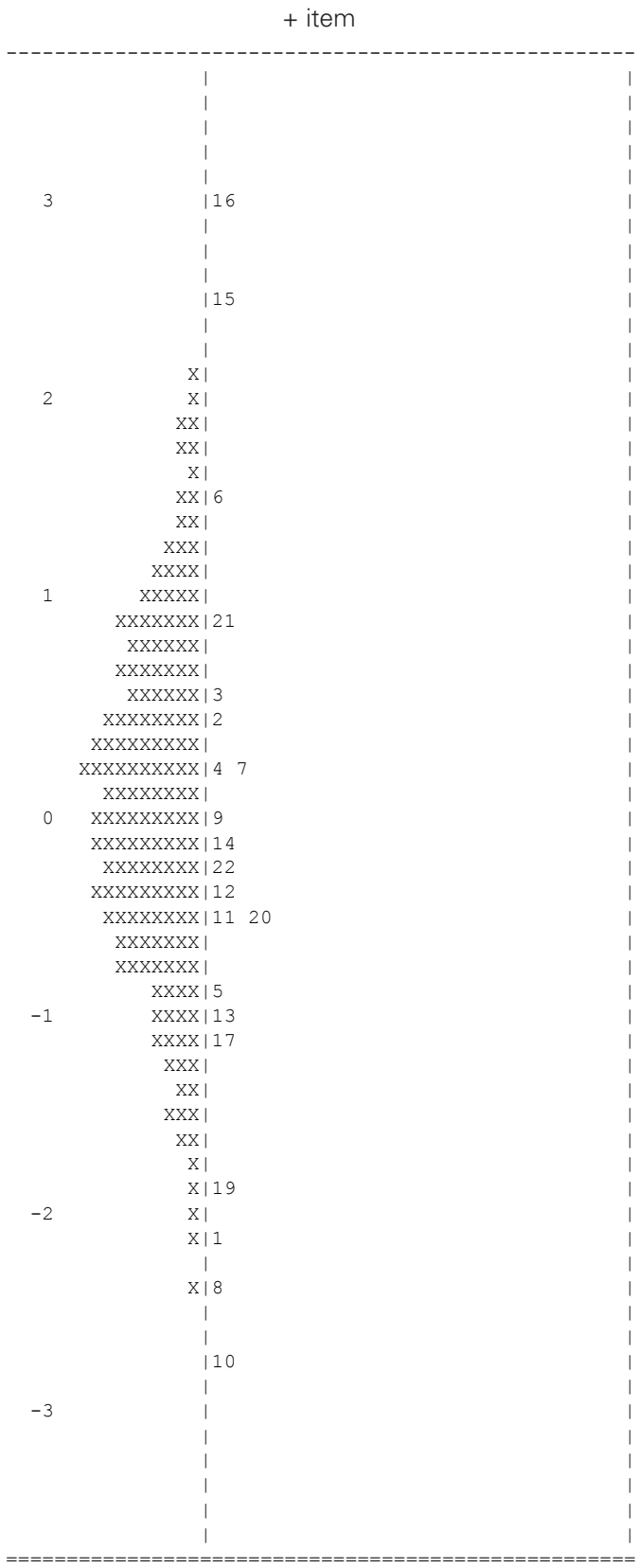


Appendix 3 Item–person maps

PAT STEM Contexts Year 3 trial calibration

Map of latent distributions and response model parameter estimates

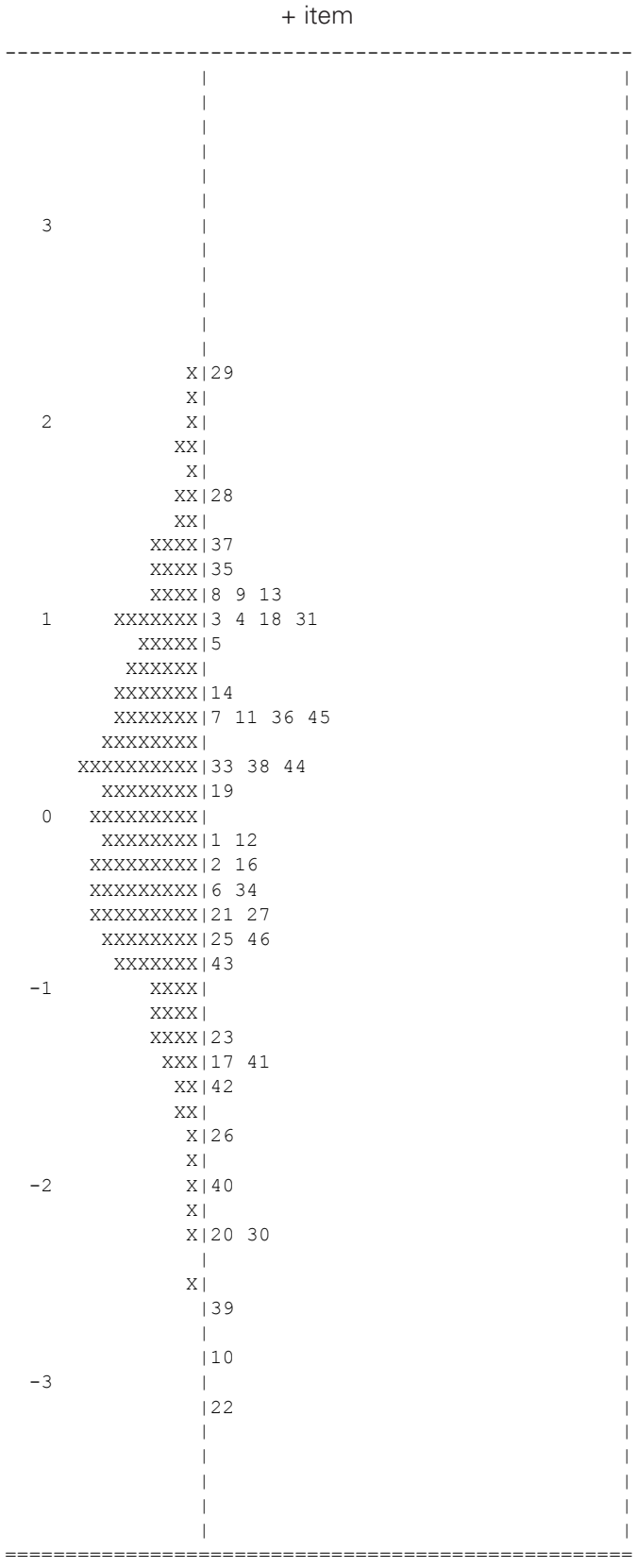
Terms in the Model (excl Step terms)



PAT STEM Contexts Year 4 trial calibration

Map of latent distributions and response model parameter estimates

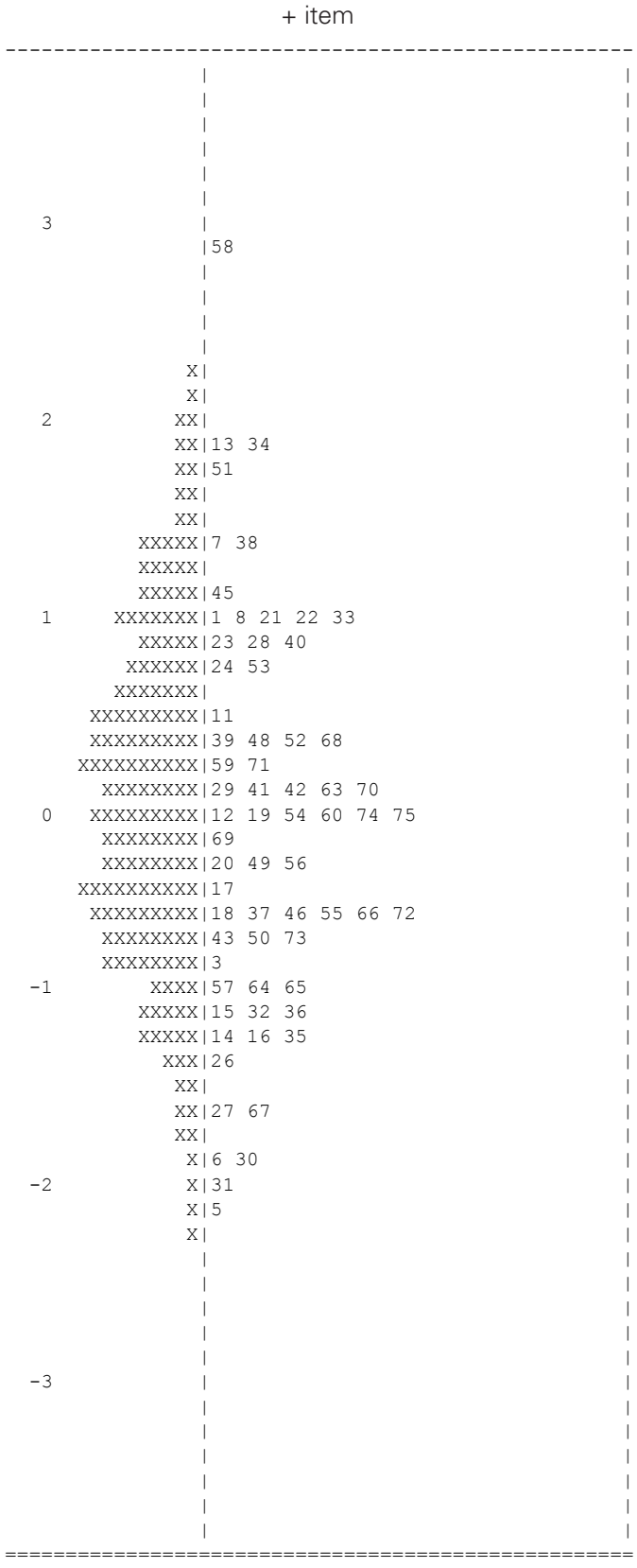
Terms in the Model (excl Step terms)



PAT STEM Contexts Year 5 trial calibration

Map of latent distributions and response model parameter estimates

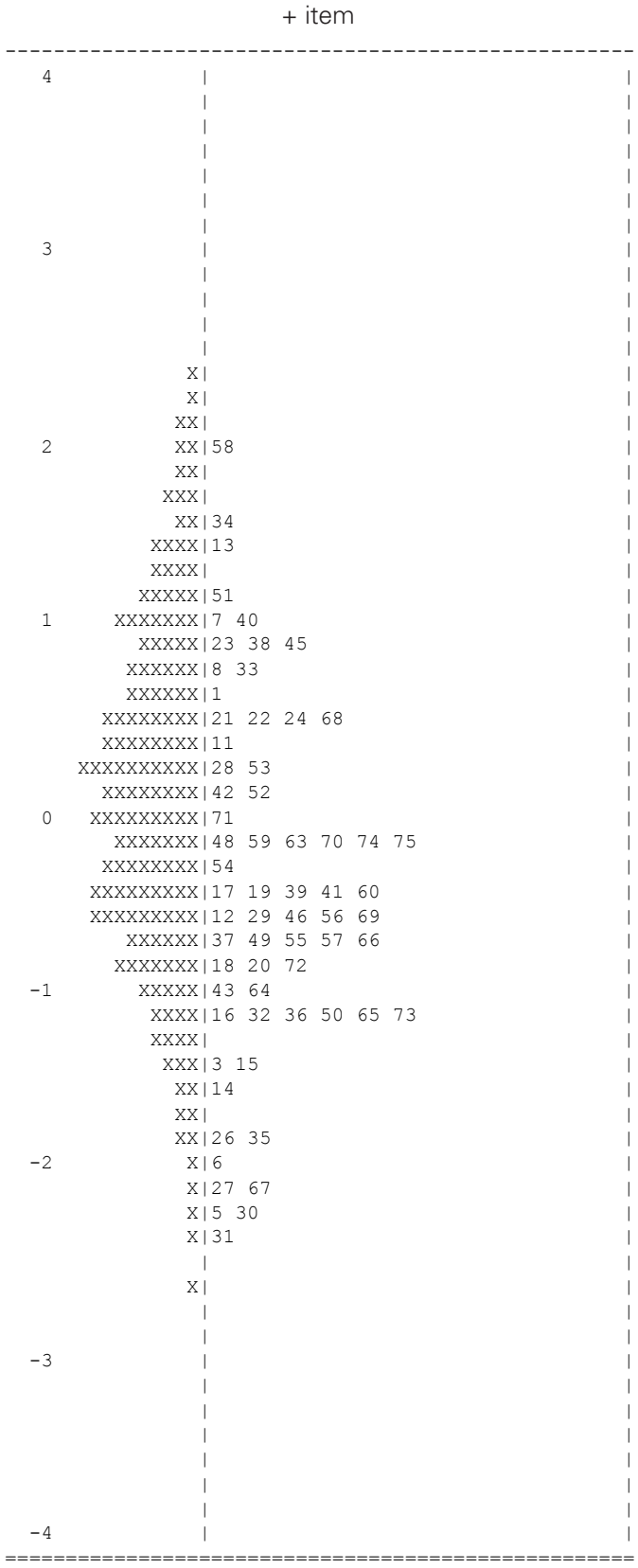
Terms in the Model (excl Step terms)



PAT STEM Contexts Year 6 trial calibration

Map of latent distributions and response model parameter estimates

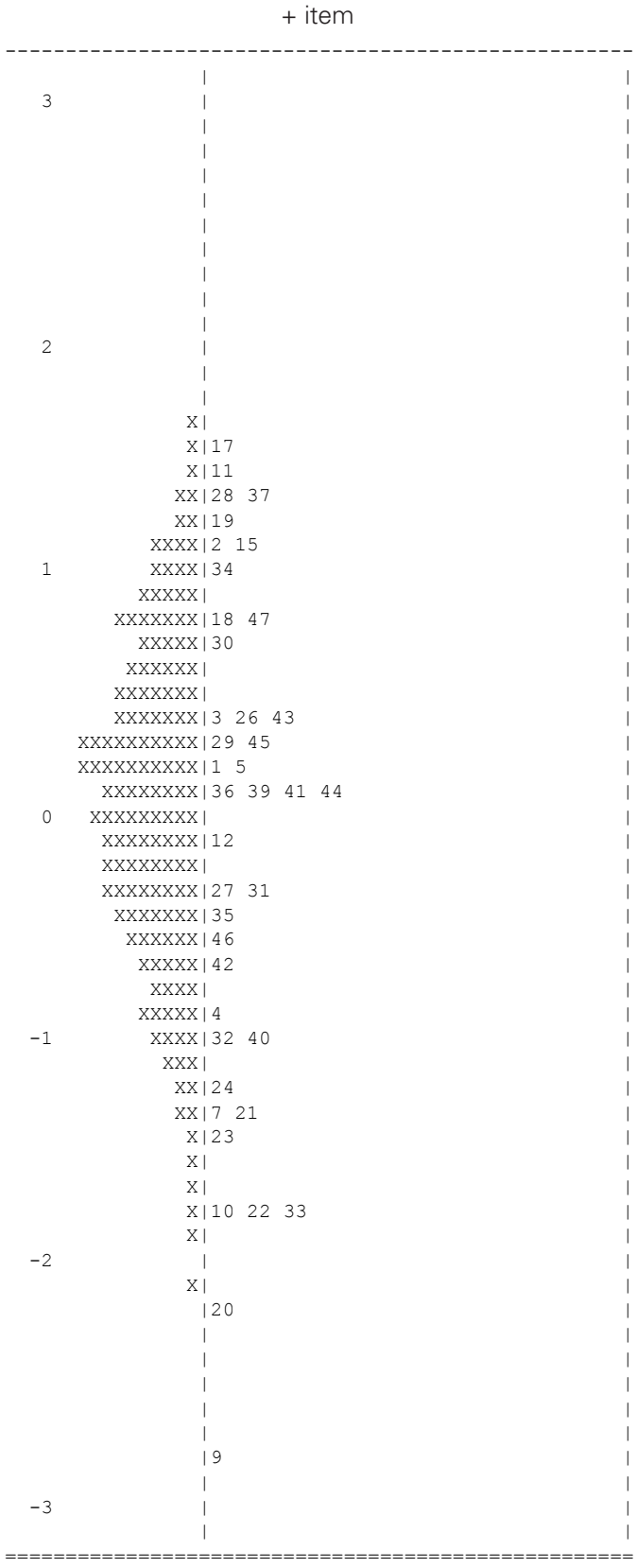
Terms in the Model (excl Step terms)



PAT STEM Contexts Year 7 trial calibration

Map of latent distributions and response model parameter estimates

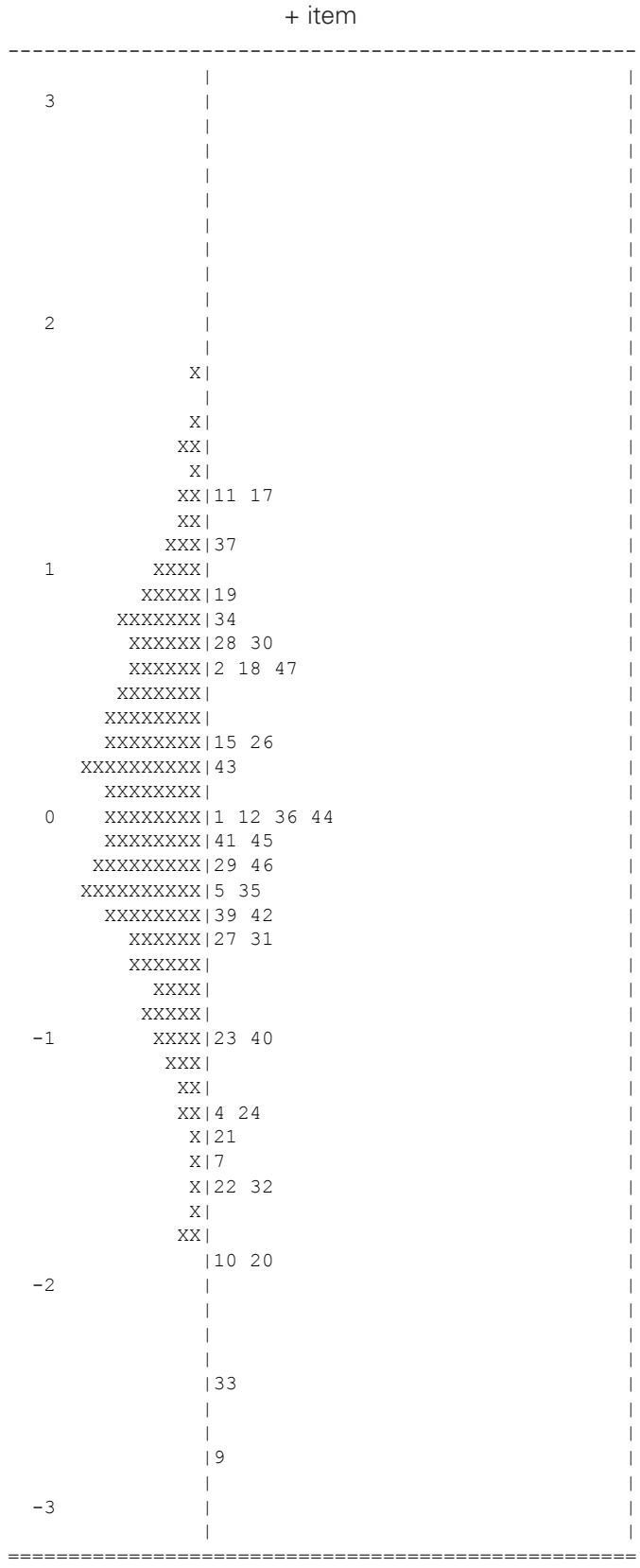
Terms in the Model (excl Step terms)



PAT STEM Contexts Year 8 trial calibration

Map of latent distributions and response model parameter estimates

Terms in the Model (excl Step terms)



Appendix 4 Scale score tranformations

The PAT STEM Contexts tests were calibrated using the Rasch model. The estimates of student ability in logits were computed based on observed raw scores in each test, and equated onto the corresponding logit scale. For reporting purposes, the estimates of ability for each test were transformed into scale scores. The unit used to express scale scores is defined from the Rasch measurement unit, the logit: 1 logit = 10 PAT STEM Contexts scale scores. This has been done to avoid assigning negative values to performance measures. The transforming formula for PAT STEM Contexts scale scores are as follows:

Scale score of item estimate = logit *10 + 120.

Scale score of ability estimate = logit *10 + 120.