Item Analysis Report

COVID-19: Monitoring Impacts on Learning Outcomes (MILO)

11 January 2021

The Global Education Monitoring (GEM) Centre supports education stakeholders to collect, analyse and use high-quality data to improve learning outcomes. The GEM Centre is a long-term partnership between the Australian Council for Educational Research (ACER) and the Australian Government's Department of Foreign Affairs and Trade (DFAT).

GPE Transforming Education

unesco Institute for Statistics

ACER

Acknowledgments

This project, the Assessment and Study of COVID Impact on Learner Progress, is referred to as the COVID-19 MILO (Monitoring Impacts on Learning Outcomes) project. This UNESCO Institute for Statistics (UIS) project is funded by the Global Partnership for Education (GPE).

The Australian Council for Educational Research (ACER) is the technical partner for this project. Support is provided from the Global Education Monitoring (GEM) Centre, an ACER initiative in partnership with the Australian government's Department of Foreign Affairs and Trade. The GEM Centre is also contributing to the UIS Global Item Bank. Technical and implementation support, and contribution to the assessment item pool, is provided by CONFEMEN.

Contents

Acknowledgments 1
Introduction 6
Background 7
The scaling model 8
Glossary 9
Item classifications 9
Data 9
Auxiliary codes 9
Response categories 10
Analysis 10
Item response model fit (weighted mean square MNSQ) 11
Discrimination coefficients 11
Item summary statistics 13
Item Map 15
Differential item functioning (DIF) 18
Itanal – item statistics 19
MM004 19
Item Characteristic Curve (ICC) 21
Results of the analysis: item treatment 21
References 23
Appendix 1a – Item descriptive statistics – MILO Mathematics 24
MM004 24
MM011 24
MM016 24
MM019 25
MM022 25
MM029 25
MM030 26
MM060 26
MM075 26
MM089 27
MM090 27

MM101 27
MM104 28
MM125 28
MM175 28
MM191 29
MM197 29
MM206 29
MM208 30
MM209 30
PM422 30
PM445 31
PM449 31
PM454 31
PM459 32
PM462 32
PM468 32
PM469 33
PM942 33
Appendix 1b – Item descriptive statistics – MILO Reading 34
MR001 34
MR002 34
MR003 34
MR024 35
MR025 35
MR033 35
MR034 36
MR035 36
MR041 36
MR042 37
MR043 37
MR044 37
MR056 37
MR058 38
MR059 38

MR069 38
MR087 39
MR089 39
MR090 39
MR201 40
MR202 40
MR203 40
MR204 41
PF449 41
PF455 41
PF456 42
PF458 42
PF487 42
PF489 43
Appendix 2a – Item Characteristic Curve (ICC) Graphs – MILO Mathematics 44
MM004 44
MM011 44
MM016 45
MM019 45
MM022 46
MM029 46
MM030 47
MM060 47
MM075 48
MM089 48
MM090 49
MM101 49
MM104 50
MM125 50
MM175 51
MM191 51
MM197 52
MM206 52
MM208 53

MM209 53
PM422 54
PM445 54
PM449 55
PM454 55
PM459 56
PM462 56
PM468 57
PM469 57
PM942 58
Appendix 2b – Item Characteristic Curve (ICC) Graphs – MILO Reading 59
MR001 59
MR002 59
MR003 60
MR024 60
MR025 61
MR033 61
MR034 62
MR035 62
MR041 63
MR042 63
MR043 64
MR044 64
MR056 65
MR058 65
MR059 66
MR069 66
MR087 67

Introduction

1) The COVID-19: Monitoring Impacts on Learning Outcomes (MILO) project aims to measure learning outcomes in six countries in Africa, in order to analyse the long-term impact of COVID-19 on learning and to evaluate the effectiveness of distance learning mechanisms utilised during school closures. In addition, this project will develop the capacity of countries to monitor learning after the crisis.

2) The four overarching goals of the project are to:

    • Evaluate the impact of COVID-19 on learning outcomes and measure the learning loss by reporting against SDG indicator 4.1.1b

    • Identify the impact of different distance learning mechanisms put in place to remediate the learning disruption generated by COVID-19

    • Expand the UIS bank of items for primary education

    • Generate a toolkit so that assessment results can be scaled to international benchmarks, reporting against SDG 4.1.1.b.

Background

3) This document describes the outcomes of applying Item Response Theory (IRT) scaling methodology to the item analysis of MILO assessment data. When applying a parametric IRT measurement model, parameter invariance is assumed. The measurement model should neither vary across sub-groups within countries nor across countries. However, it needs to be recognised that for most items there will always be a certain amount of item-by-country interaction effects and that complete parameter invariance cannot be expected given the difference in cultures and languages that are included in this international study.

4) To investigate the psychometric properties of items and parameter invariance cognitive items were calibrated at national level first. Resulting item statistics were compared and items that exhibited problems were flagged for further review by item developers. The outcomes of this review were used to make a decision about how to treat each item in each country. A few items were excluded from MILO scaling altogether as they have poor psychometric characteristics in all countries and some items were deleted from the scaling in particular countries as they have poor psychometric characteristics in those particular countries but function well in the others.

5) Details of item treatments are described in the section of this report titled **Results of the analysis: item treatment.**

6) This document provides a brief description of the IRT scaling model. This is followed by a glossary of the psychometric terminology that is relevant to the item summary table, item maps and the item statistics in a tabular form (itanal) for the final international item calibrations by domain at the end of the process outlined above. Terminology that is confined to a particular report is also explained in the relevant section of this document where required.

The scaling model

8) Cognitive items were scaled using item response theory (IRT) scaling methodology. The Mixed Coefficients Multinomial Logit Model (MCMLM) as described by Adams, Wilson and Wang (1997) was used to scale the MILO data. Psychometric analysis included item level analysis (item calibration at national and international level) and proficiency level generation.

9) The analysis of item characteristics and the estimation of model parameters were carried out with the ACER ConQuest® software package (Version 5 software: see Adams, Wu, Macaskill, Haldane, Sun & Cloney, 2021).

10) The use of the MCMLM model which a generalised form of the Rasch model (Rasch, 1960) was proposed because:

    - Of all available item response theory models, it provides the strictest assessment of psychometric validity.
    - It supports the construction and validation of meaningful described proficiency scales.
    - The Rasch family as implemented in ACER ConQuest® can be, and is, used to explore and control item position effects. Further, the model can be routinely applied in contexts that require multidimensional scaling.
    - It supports equating tests for the purposes of maintaining and monitoring the validity of the historic assessments.

11) Please refer to Appendix B of the MILO main report regarding details of the scaling procedures.

Glossary

Item classifications

12) The two domains of MILO 2021 are mathematics and reading. The items in the reports are reported separately by domain. Reports for each domain are provided only at international level.

13) The analysis distinguishes between two types of items: those that are **dichotomous** and those that are **partial credit**.

14) **Dichotomous items** are those for which an incorrect response obtains a score of 0 and a correct response obtains a score of 1. Multiple-choice items with a maximum score of 1 are classified as dichotomous items.

15) **Partial credit items** are usually those for which an incorrect response obtains a score of 0, a partially correct response obtains a score of 1, and a fully correct response obtains a score of 2. In MILO 2021, there are two Mathematics items scored as partial credit items.

Data

16) Student responses are first coded by data entry operators or domain expert coders. The codes applied at this stage are called the raw codes. The raw codes can be valid or auxiliary codes. Some of the raw codes are re-coded during data cleaning.

17) **Valid codes** are the ones that correspond to the student responses (if they can be entered into the database) or the coder responses. Examples of valid codes are code 1 for option A circled by a student in a multiple choice item and code 0 circled by a coder in the constructed response manual item. All valid codes are included in the item analysis.

18) No items were marked by domain expert coders in MILO 2021.

Auxiliary codes

19) A **Missing** value (9, 99,…) is assigned if a student did not answer an item. The missing values embedded between valid codes are differentiated from the missing values clustered at the end of the booklet. During data cleaning all consecutive missing values clustered at the end of each booklet, except for the first missing value of the series, are replaced by a Not Reached code (r). Missing values (embedded) are included when scaling the data and scored 0. Not reached missing values are excluded from the item analysis.

20) An **Invalid** value (8, 98,…) is assigned if a student response could not be entered into the database (e.g. if two choices are circled in a multiple choice item). Invalid values are treated as code 0 during item analysis.

21) An **N/A** value (7, 97,…) is assigned if an item was not administered to a student. N/A values are excluded from the item analysis.

22) The **number of valid cases** is equal to the number of valid codes for the item in the database. All statistics presented should be treated with caution if the number of valid cases per category is less than fifteen.

Response categories

23) **Response categories** are the codes that are used in item analysis. Information about which response categories are equal to raw codes and which response categories are equal to recoded raw codes for different item formats is given below:

- The response categories of all multiple choice items have the same raw codes according to the number of choices presented in each item. Some of the raw missing codes are equal to response category r ('not reached').
- Two Mathematics items are complex multiple choice items. The raw codes for N/A, invalid and missing are equal to response categories 9997, 9998 and 9999 respectively. For the purpose of analysis, they were recoded to a single-digit code 7, 8 or 9 accordingly and in some cases equal to response category r ('not reached'). For the remaining raw valid codes, the response categories are assigned so that they reflect the number of elements in the complex multiple choice that are correct.

Analysis

24) **Item Response Theory (IRT)** is a methodology that is used to analyse items in MILO 2021. Items were analysed using ConQuest software. Information from the ConQuest item analysis output files is included in the reports. The following terminology related to the IRT statistics is used throughout the reports:

- **Item difficulty (delta)** in the IRT models is defined as the level of ability at which the probability of success on the item (or partial success for the **partial credit items**) is 0.5. Item difficulty is measured in logits[1].
- **Logit scale** is a log-linear transformation of the probability scale[2]. Probability *p=0.5* is transformed into 0, probability of approximately 0.76 ($p = \frac{e}{1+e}$) is transformed into 1, probability of approximately 0.91 ($p = \frac{e^2}{1+e^2}$) is transformed into 2, and so on.
- **Item threshold** for a score category for an item is defined as the ability at which the probability of achieving that score or higher reaches 0.5.
- **Adjusted domain score** is a student's total score for a domain minus the student's score for the particular item.

[1] Please note that in the IRT models item difficulty and student abilities are measured simultaneously.
[2] The transformation is $ln\frac{p}{1-p}$.

- **Item-rest correlation** is a correlation between students' scores on an item and their **adjusted domain scores**.
- **Fit** or **weighted MNSQ** is the Weighted Infit Mean Square statistic (Wu, 1997). This statistic compares (by the analysis of residuals) predicted scores for an item and the students' observed scores for the item. Please refer to section "Item response model fit" below for further details.
- **Point-biserial (PB) correlation** is a correlation between a response category re-coded as a dummy variable (a code of 1 for students with correct response category and a code of 0 for students with all other response categories) and **adjusted domain scores**. Please refer to section "Discrimination coefficients" below for further details.

25) When reviewing the item calibration results, particular attention was paid to the fit of the items to the scaling model and item discrimination.

Item response model fit (weighted mean square MNSQ)

26) For each item parameter, the ACER ConQuest® fit mean square index is used to provide an indication of the compatibility of the model and the data. For each student, the model describes the probability of obtaining the different item scores. It is therefore possible to compare the model prediction and what has been observed for one item across students. Accumulating comparisons across students gives an item-fit statistic. As the fit statistics compare an observed value with a predicted value, the fit is an analysis of residuals. In the case of the item weighted MNSQ fit statistics, values near one are desirable. A weighted MNSQ greater than one is associated with a low discrimination index, meaning the data exhibit more variability than expected by the model, and an infit mean square less than one is associated with a high discrimination index, meaning the data exhibits less variability than expected by the model.

Discrimination coefficients

27) For each item, the correlation between the students' score and aggregate score on the set for the same domain as the item of interest is used as an index of discrimination. If p_{ij} (calculated as x_{ij}/m_j)[3] is the proportion of score levels that student i achieved on item j, and $p_i = \sum_j p_{ij}$ (where the summation is of the items from the same booklet and domain as item j) is the sum of the proportions of the maximum score achieved by student i, then the discrimination is calculated as the product-moment correlation between p_{ij} and p_i for all students. For dichotomous items such as multiple-choice items, this index will be the usual point-biserial index of discrimination.

28) The point-biserial index of discrimination for a particular category of an item is a comparison of the aggregate score between students selecting that category and all

[3] Where m_j is the maximum score for an item j, and x_{ij} is the student i score for item j.

other students. If the category is the correct answer, the point-biserial index of discrimination should be higher than 0.20 (Ebel and Frisbie, 1986). They set out the following recommendations regarding the index of discrimination shown in Table 1.

Table 1: Point-Biserial Index Ranges

Magnitude	Comment	Recommended action for item
> 0.39	Excellent	Retain
0.30 – 0.39	Good	Possibilities for improvement
0.20 – 0.29	Mediocre	Need to check/review
0.00 – 0.20	Poor	Discard or review in depth
< -0.01	Worst	Definitely discard

29) Non-key categories (i.e. incorrect options or 'distractors') should have a negative point-biserial index of discrimination. The point-biserial index of discrimination for a partial credit item should be ordered, i.e. categories scored 0 should have a lower point-biserial correlation than the categories scored 1, and so on.

Item summary statistics

30) A table of item summary statistics was generated for each domain. Table 2 and Table 3 are the tables of item summary statistics of MILO Mathematics and Reading, respectively.

31) Item summary statistics include the following information:

- Facility– percent correct with countries equally weighted
- Item-Rest Correlation (Item-Rest Cor). – equal to point biserial of a correct category response
- Weighted MNSQ (Wghtd MNSQ) – item fit
- Delta – item difficulty in logits, with a mean of 0 and standard deviation of 1.
- Gender DIF - Differential item functioning by gender (Please refer to section "Differential item functioning" for details)

Table 2: Item summary statistics – MILO Mathematics

Item	Student count	Facility*	Item-Rest Cor	Wghtd MNSQ	Delta	Gender DIF
M_MM004	28149	74%	0.45	0.88	-1.89	-0.08
M_MM011	27588	43%	0.41	0.95	-0.33	0.02
M_MM016	28149	38%	0.34	1.02	-0.05	0.01
M_MM019	28136	33%	0.44	0.93	0.22	-0.03
M_MM022	28150	50%	0.38	1.00	-0.63	-0.01
M_MM029	27710	35%	0.23	1.11	0.09	0.07
M_MM030	10055	34%	0.34	1.48	0.31	0.04
M_MM060	28147	59%	0.38	1.00	-1.10	-0.02
M_MM075	28075	19%	0.14	1.11	1.06	-0.03
M_MM089	27896	31%	0.20	1.15	0.28	0.00
M_MM090	27663	22%	0.34	0.99	0.82	-0.03
M_MM101	10098	31%	0.31	1.13	0.59	-0.02
M_MM104	10003	31%	0.33	1.11	0.61	-0.01
M_MM125	27534	53%	0.48	0.90	-0.85	0.05
M_MM175	28087	18%	0.24	1.06	1.19	0.02
M_MM191	28053	18%	0.28	1.03	1.17	0.13
M_MM197	27961	17%	0.27	1.03	1.21	0.01
M_MM206	27960	6%	0.25	0.97	2.48	0.25
M_MM208	28126	52%	0.50	0.89	-0.74	0.03
M_MM209	27825	37%	0.25	1.26	0.02	0.03
M_PM422	28149	69%	0.49	0.87	-1.59	-0.04
M_PM445	27788	45%	0.51	0.87	-0.44	-0.04
M_PM449	28018	31%	0.26	1.10	0.30	-0.02
M_PM454	28067	29%	0.44	0.92	0.41	-0.04
M_PM459	27969	54%	0.49	0.89	-0.83	0.01
M_PM462	28062	65%	0.50	0.86	-1.40	0.02
M_PM468	28116	48%	0.49	0.91	-0.57	-0.01
M_PM469	28055	38%	0.50	0.90	-0.08	0.00
M_PM942	28092	43%	0.36	1.00	-0.28	-0.05

*Note: Facility, percentages of correct responses, was computed with countries equally weighted.

Table 3: Item summary statistics – MILO Reading

Item	Student count	Facility*	Item-Rest Cor	Wghtd MNSQ	Avg Delta	Gender DIF
R_MR001	23661	72%	0.52	0.83	-1.45	-0.03
R_MR002	23655	68%	0.56	0.80	-1.18	0.04
R_MR003	23654	67%	0.55	0.83	-1.10	0.01
R_MR024	23623	28%	0.49	0.91	0.99	0.02
R_MR025	23616	53%	0.46	0.96	-0.39	0.01
R_MR033	13449	32%	0.24	1.12	0.58	0.03
R_MR034	10068	26%	0.34	1.12	1.36	0.10
R_MR035	23462	35%	0.42	1.01	0.57	-0.04
R_MR041	23405	30%	0.34	1.08	0.83	-0.02
R_MR042	23390	26%	0.35	1.05	1.07	-0.02
R_MR043	23650	68%	0.42	0.96	-1.19	-0.01
R_MR044	23627	58%	0.35	1.08	-0.64	-0.05
R_MR056	23613	71%	0.55	0.80	-1.35	-0.02
R_MR058	23601	56%	0.59	0.82	-0.54	0.00
R_MR059	23595	49%	0.38	1.06	-0.16	-0.03
R_MR069	23382	52%	0.42	1.00	-0.34	-0.06
R_MR087	23560	34%	0.48	0.93	0.65	0.01
R_MR089	23543	35%	0.51	0.92	0.59	-0.03
R_MR090	23532	42%	0.45	0.99	0.16	-0.01
R_MR201	23592	45%	0.38	1.06	0.02	0.01
R_MR202	23578	35%	0.44	0.98	0.58	0.02
R_MR203	23573	47%	0.32	1.13	-0.10	0.06
R_MR204	23566	30%	0.21	1.22	0.83	0.03
R_PF449	23664	75%	0.36	1.00	-1.60	0.03
R_PF455	23261	38%	0.53	0.90	0.38	-0.06
R_PF456	23210	32%	0.32	1.10	0.69	-0.01
R_PF458	13143	23%	0.29	1.03	1.13	0.03
R_PF487	23663	60%	0.35	1.08	-0.76	-0.05
R_PF489	23662	39%	0.37	1.07	0.35	0.00

*Note: Facility, percentages of correct responses, was computed with countries equally weighted.

Item Map

32) Figure 1 and Figure 2 presents item maps of MILO Mathematics and Reading, respectively. The crosses represent students, the numbers represent items, and in the case of a partial credit item the threshold is included (for example item 21.1 represents item 21.1, score = 1, partial credit; and 21.2 represents item 21, score = 2, full credit). The vertical line represents the measured domain scale with high-performing students and difficult items at the top and low-performing students and easy items at the bottom.

33) The two scales are not directly comparable because they are different latent traits and have been calibrated separately.

34) The response probability in these figures is 0.5, which means that students with an ability equal to the difficulty (or threshold) of an item have a 50 per cent chance of responding correctly to that item.

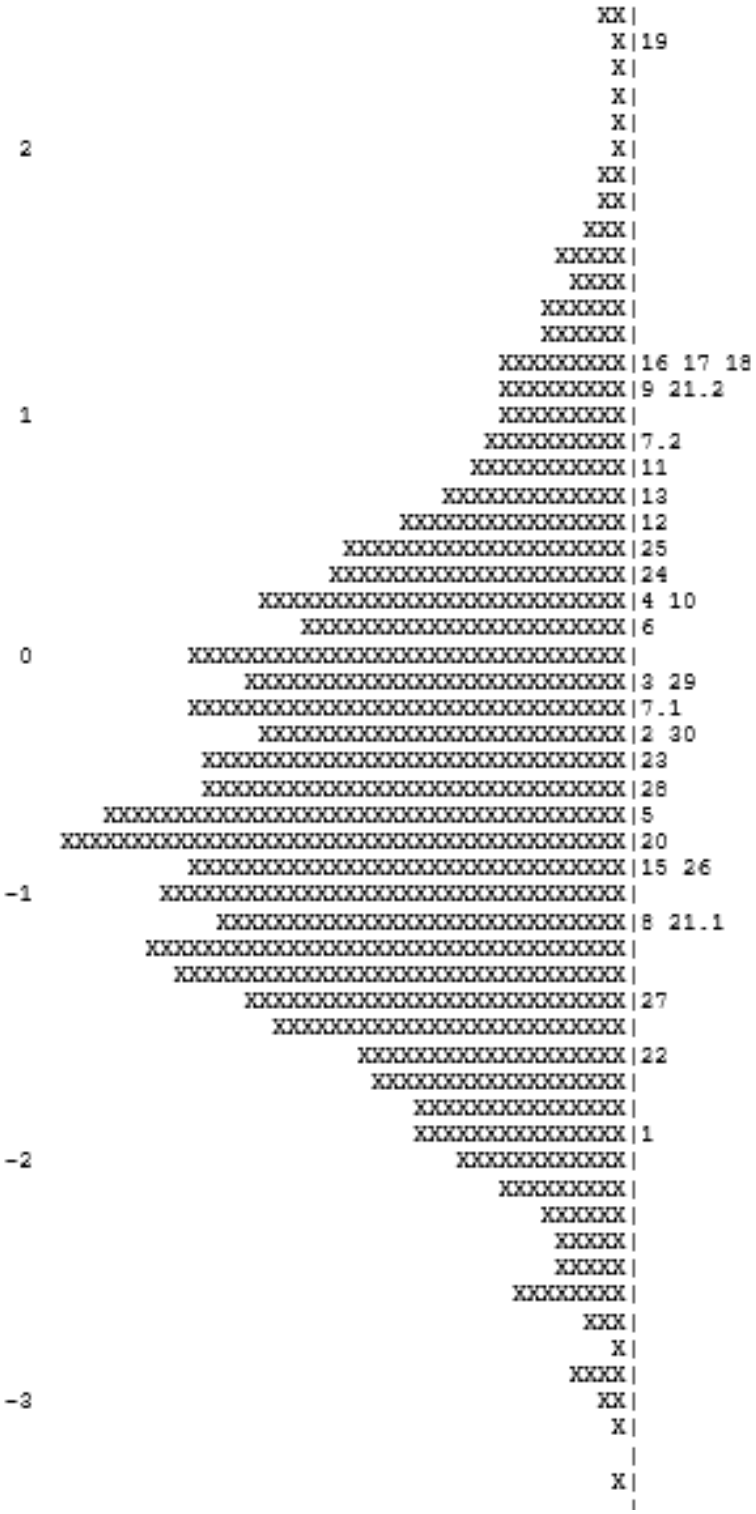


Figure 1: Item maps for MILO Mathematics

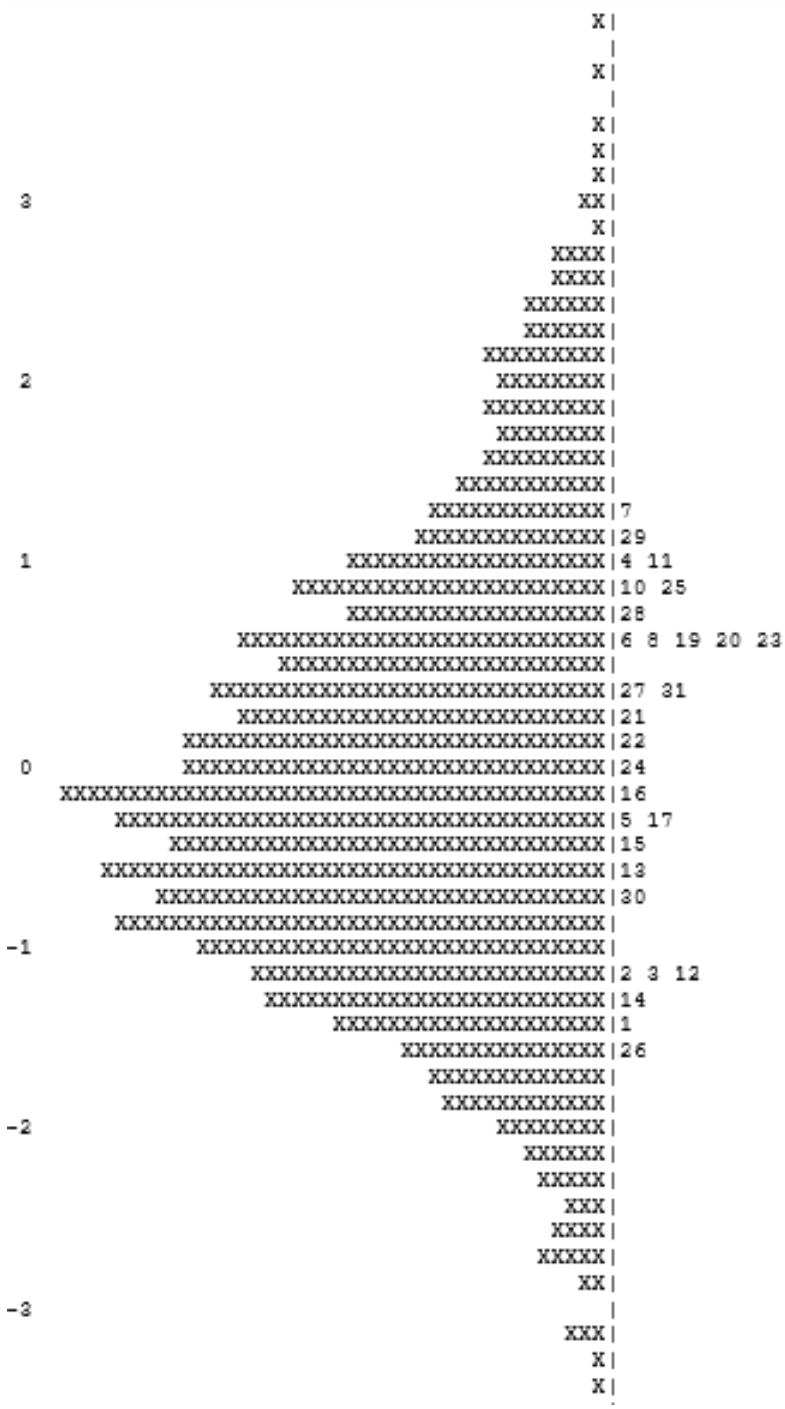


Figure 2: Item maps for MILO Reading

Differential item functioning (DIF)

35) The quality of the items was also explored by assessing differential item functioning (DIF) by gender for each country and domain. DIF occurs when groups of students with the same ability have different probabilities of responding correctly to an item.

36) For example, if a group of boys with the same average ability as a group of girls have a higher probability of success for a particular item, that item shows DIF in favour of boys. This constitutes a violation of the model, which assumes that the probability is only a function of ability (and item difficulty) and not of any other variable.

37) Substantial item DIF (e.g. < -0.3 or > 0.3[4]) with respect to gender may result in bias of performance estimates across gender groups.

38) The gender DIF estimates range between -0.084 and 0.246 for MILO Mathematics and between -0.063 and 0.104 for MILO Reading. No instances of substantial gender DIF were encountered so no items were removed for this reason.

39) Gender DIF estimates are in Table 2 and Table 3 in section "Item summary statistics".

[4] Magnitude of item by gender interaction estimates from a facet model. See PISA 2006 Technical Report (OECD, 2009a).

Itanal – item statistics

41) An item analysis (itanal) of item descriptive statistics by item in tabular form was produced for each domain.

Table 4: Example of Item Statistics in Tabular Form

MM004

Response	Score	Count	Per cent	Pt Bis	Mean Ability	SD Ability
1	0	1702	6.0	-0.16	-1.21	0.73
2	**1**	**21205**	**75.2**	**0.45**	**-0.27**	**0.93**
3	0	1784	6.3	-0.14	-1.15	0.83
4	0	2273	8.1	-0.24	-1.44	0.76
8	0	764	2.7	-0.25	-2.39	1.05
9	0	421	1.5	-0.12	-1.74	1.21
r	0	58	0.2	-0.09	-3.01	0.94

42) Table 4 is extracted from the itanal of MILO Mathematics, showing the item statistics of test item M_MM004. The tabular item statistics include the following information:

- Score – shows score assigned to each response category (response 2 is correct response- scored 1)
- Per cent – percentage of the responses given in each category by all students that answered the question
- Point Biserial correlation - percentages for each category, category correlations with overall score (Pt Bis)
- Mean Ability - the average ability within each category

43) Item MM004 was done by 28,207 students in MILO Mathematics.

44) The first column Response are the original response categories. This is a multiple-choice item and therefore, the responses are: 1=A, 2=B, 3=C, 4=D, 8='multiple response', 9='missing' and r='not reached'. The second column shows the score assigned to each response category. The correct response to this item is 2 (B).

45) The third and fourth columns in Table 4 list the number and percentage of students in each category. For example, 21,205 students (75.2%) gave the correct response.

46) The point-biserial correlations (Pt Bis) are presented in column five. This is the correlation between a response category coded as a dummy variable (a score of 1 for students that responded with the current code and a score of 0 for students in other response categories) and the adjusted domain score. Correct responses should have

positive correlations with the total score, incorrect responses negative correlations. In this case all of the incorrect responses have negative point-biserials and the key category has a positive point-biserial of 0.45.

47) The two last columns (Mean Ability and SD Ability) show the average ability of students responding in each category and the associated standard deviation. If an item is functioning well the group of students that gave the correct response should have a higher mean ability than the groups of students that provided each of the incorrect responses. This is true for this item.

48) Appendices 1a and 1b are the itanals of item statistics of MILO Mathematics and Reading, respectively. Both appendices have response category 'r' (not reached) included.

Item Characteristic Curve (ICC)

49) For each item, an Item Characteristic Curve (ICC) graph was produced. Figure 3 is an ICC graph by category for test item MM004.

50) With increasing ability (latent trait in logits) the probability of a correct response increases. The solid line shows the expected item responses given the IRT model and the dot points indicate the average percentage of correct responses in different ability groups.

51) If the slope for the observed correct response is shallower than the expected curve, the item is less discriminating than predicted by the model; if it is steeper the item has a higher discrimination than expected.

52) Key is flagged with "*" in the legend.

53) In Figure 3, the test item shows close to perfect fit to the scaling model.

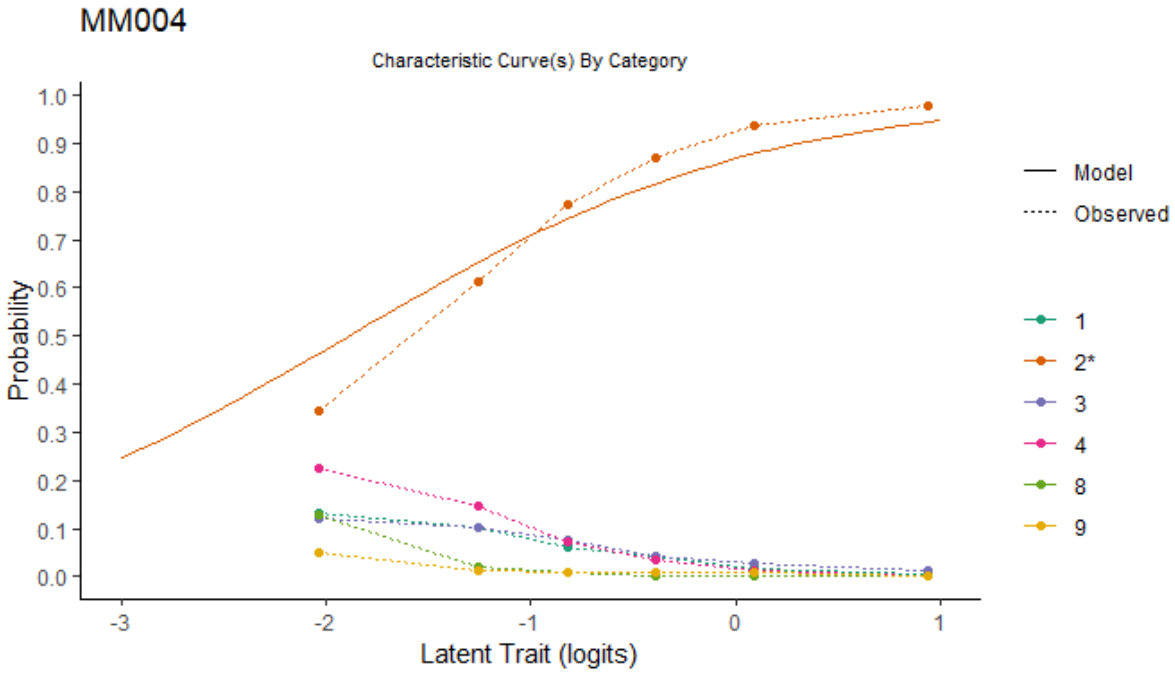


Figure 3: Item Characteristic Curve (ICC) by category

54) Appendices 2a and 2b are the ICC graphs of MILO Mathematics and Reading items respectively.

Results of the analysis: item treatment

55) The psychometric properties of the reading items administered in Burundi was unexpectedly inconsistent with those of the other countries. In particular, the response patterns in nearly all of the reading items was consistent with high rates of guessing and resulted in very low discrimination. It was therefore decided to exclude them from the international reading item calibration. Burundi student

reading proficiency estimations were subsequently based on the international calibration.

56) After examining the proportion of responses in each item response category and the ordering of mean abilities of students across response categories, two Mathematics items (M_MM030 and M_MM209) were re-scored as partial credit items with a maximum score of two. That is, the response previously marked as correct was scored as 2, and another response selected by able students was given a partial credit of a score of 1 rather than being marked as incorrect.

57) Final decisions on removing test items were based on a range of different criteria. Generally, items were flagged for review if initial item calibrations showed a considerably higher infit statistic (e.g. MNSQ > 1.2) as well as low item-rest correlation (0.2 or lower). The ACER project team considered both item fit criteria as well as the content of the item prior to a decision about removing or retaining flagged items for scaling.

58) Of the 30 Mathematics and 31 Reading items in the original test, one Mathematics (M_MM109) and two Reading items (R_MR040 and R_MR070) were removed from the scale due to poor fit statistics at international level. Consequently, these items were not used to estimate student achievement.

59) In addition, three Mathematics items (M_MM030, M_MM101 and M_MM104) and one Reading item (R_MR034) were removed for French. Another two Reading items (R_MR033 and R_PF458) were removed for English. Consequently, these items were not used to estimate student achievement in the respective language groups.

60) After examining the item and test level statistics, and excluding some poor performing items, the mathematics test contained 26 items for French-based assessments and 29 items for English-based assessments. The reading assessments were contained 28 items for French-based assessments and 27 items for English-based assessments.

References

Adams, R. J., Wu, M. L., Macaskill, G., Haldane, S. A., Sun, X. X. & Cloney, D. (2021). *ACER ConQuest Version 5: Generalised item response modelling software [Computer software]*. Melbourne: ACER.

Ebel, R.L. and D.A. Frisbie. (1986). *Essentials of Education Measurement*, Prentice Hall, Englewood Cliffs, New Jersey.

Masters, G. N., & Wright, B. D. (1997). 'The partial credit model.' In W.J. Van der Linden & R.K. Hambleton (eds), *Handbook of Modern Item Response Theory*, pp. 101–122. New York/Berlin/Heidelberg: Springer.

Rasch, G. (1960). *Probabilistic Models for Some Intelligence and Attainment Tests*. Copenhagen: Nielsen and Lydiche.

Wu, M. L. (1997). *The development and application of a fit test for use with marginal maximum likelihood estimation and generalised item response models* [M.Ed, The University of Melbourne]. http://hdl.handle.net/11343/57769

Appendix 1a – Item descriptive statistics – MILO Mathematics

MM004

Response	Score	Count	Per cent	Pt Bis	Mean Ability	SD Ability
1	0	1702	6.0	-0.16	-1.21	0.73
2	**1**	**21205**	**75.2**	**0.45**	**-0.27**	**0.93**
3	0	1784	6.3	-0.14	-1.15	0.83
4	0	2273	8.1	-0.24	-1.44	0.76
8	0	764	2.7	-0.25	-2.39	1.05
9	0	421	1.5	-0.12	-1.74	1.21
r	0	58	0.2	-0.09	-3.01	0.94

MM011

Response	Score	Count	Per cent	Pt Bis	Mean Ability	SD Ability
1	0	5120	18.2	-0.06	-0.74	0.83
2	**1**	**12693**	**45.0**	**0.42**	**-0.04**	**0.96**
3	0	3163	11.2	-0.09	-0.85	0.80
4	0	4102	14.5	-0.19	-1.04	0.85
8	0	731	2.6	-0.20	-2.06	1.21
9	0	1779	6.3	-0.15	-1.23	1.01
r	0	615	2.2	-0.14	-1.65	1.13

MM016

Response	Score	Count	Per cent	Pt Bis	Mean Ability	SD Ability
1	0	6048	21.4	-0.12	-0.81	0.88
2	**1**	**11127**	**39.4**	**0.34**	**-0.06**	**1.01**
3	0	4194	14.9	-0.08	-0.78	0.93
4	0	2110	7.5	-0.09	-0.89	0.83
8	0	3089	11.0	-0.14	-1.07	1.12
9	0	1601	5.7	-0.10	-1.05	1.09
r	0	37	0.1	-0.07	-3.10	1.08

MM019

Response	Score	Count	Per cent	Pt Bis	Mean Ability	SD Ability
1	0	2216	7.9	-0.14	-1.04	0.85
2	0	3765	13.4	-0.13	-0.88	0.82
3	0	9918	35.2	-0.07	-0.70	0.83
4	**1**	**9641**	**34.2**	**0.44**	**0.11**	**0.96**
8	0	1362	4.8	-0.26	-1.84	1.09
9	0	1234	4.4	-0.16	-1.38	1.05
r	0	62	0.2	-0.10	-2.97	0.94

MM022

Response	Score	Count	Per cent	Pt Bis	Mean Ability	SD Ability
1	0	4893	17.3	-0.07	-0.77	0.87
2	0	3696	13.1	-0.07	-0.79	0.86
3	**1**	**14324**	**50.8**	**0.38**	**-0.13**	**0.95**
4	0	3333	11.8	-0.22	-1.18	0.87
8	0	1076	3.8	-0.23	-1.90	1.18
9	0	828	2.9	-0.14	-1.49	1.09
r	0	57	0.2	-0.09	-3.02	0.94

MM029

Response	Score	Count	Per cent	Pt Bis	Mean Ability	SD Ability
1	**1**	**10064**	**35.7**	**0.25**	**-0.13**	**0.98**
2	0	2819	10.0	-0.04	-0.73	0.93
3	0	7480	26.5	0.04	-0.55	0.92
4	0	5183	18.4	-0.09	-0.80	0.96
8	0	768	2.7	-0.16	-1.75	1.34
9	0	1396	5.0	-0.18	-1.45	1.03
r	0	492	1.7	-0.15	-1.85	1.10

MM030

Response	Score	Count	Per cent	Pt Bis	Mean Ability	SD Ability
1	0	3652	36.1	-0.19	-0.66	0.94
2	1	2534	25.1	0.13	-0.01	1.04
3	**2**	**2321**	**23.0**	**0.26**	**0.38**	**1.12**
4	0	1404	13.9	-0.16	-0.85	1.11
8	0	36	0.4	-0.05	-1.33	1.27
9	0	108	1.1	-0.10	-1.39	1.07
r	0	58	0.6	-0.11	-2.19	1.27

MM060

Response	Score	Count	Per cent	Pt Bis	Mean Ability	SD Ability
1	0	5025	17.8	-0.01	-0.66	0.94
2	**1**	**16866**	**59.8**	**0.38**	**-0.20**	**0.92**
3	0	2743	9.7	-0.20	-1.19	0.78
4	0	2064	7.3	-0.22	-1.39	0.83
8	0	791	2.8	-0.25	-2.29	1.08
9	0	658	2.3	-0.14	-1.59	1.10
r	0	60	0.2	-0.09	-2.98	0.95

MM075

Response	Score	Count	Per cent	Pt Bis	Mean Ability	SD Ability
1	0	9274	32.9	0.14	-0.40	0.92
2	0	6849	24.3	-0.02	-0.60	0.88
3	0	3146	11.2	-0.07	-0.77	0.87
4	**1**	**5521**	**19.6**	**0.15**	**-0.14**	**1.16**
8	0	1188	4.2	-0.24	-1.88	1.16
9	0	2097	7.4	-0.14	-1.10	1.09
r	0	124	0.4	-0.12	-2.75	1.05

MM089

Response	Score	Count	Per cent	Pt Bis	Mean Ability	SD Ability
1	0	4530	16.1	0.09	-0.40	0.98
2	**1**	**9006**	**31.9**	**0.20**	**-0.16**	**1.00**
3	0	7119	25.2	-0.06	-0.69	0.91
4	0	4900	17.4	-0.03	-0.65	0.90
8	0	726	2.6	-0.20	-2.08	1.21
9	0	1615	5.7	-0.20	-1.45	1.05
r	0	308	1.1	-0.14	-2.14	1.07

MM090

Response	Score	Count	Per cent	Pt Bis	Mean Ability	SD Ability
1	0	3466	12.3	-0.02	-0.63	0.88
2	**1**	**6694**	**23.7**	**0.35**	**0.15**	**1.02**
3	0	3075	10.9	-0.08	-0.79	0.90
4	0	12359	43.8	-0.06	-0.66	0.89
8	0	491	1.7	-0.19	-2.27	1.23
9	0	1578	5.6	-0.17	-1.32	1.05
r	0	538	1.9	-0.15	-1.78	1.10

MM101

Response	Score	Count	Per cent	Pt Bis	Mean Ability	SD Ability
1	0	1997	19.7	-0.01	-0.35	1.04
2	0	1812	17.9	-0.05	-0.46	1.02
3	**1**	**3301**	**32.6**	**0.31**	**0.27**	**1.12**
4	0	2826	27.9	-0.25	-0.77	1.01
8	0	34	0.3	-0.07	-1.83	1.50
9	0	128	1.3	-0.07	-0.94	1.31
r	0	15	0.1	-0.07	-3.04	1.27

MM104

Response	Score	Count	Per cent	Pt Bis	Mean Ability	SD Ability
1	0	2148	21.2	-0.09	-0.53	1.00
2	0	1855	18.3	-0.13	-0.62	0.99
3	**1**	**3245**	**32.1**	**0.33**	**0.31**	**1.12**
4	0	2603	25.7	-0.11	-0.55	1.07
8	0	35	0.3	-0.04	-1.23	1.40
9	0	117	1.2	-0.08	-1.11	1.13
r	0	110	1.1	-0.10	-1.54	1.33

MM125

Response	Score	Count	Per cent	Pt Bis	Mean Ability	SD Ability
1	0	3537	12.5	-0.18	-1.05	0.79
2	**1**	**15480**	**54.9**	**0.48**	**-0.08**	**0.93**
3	0	2189	7.8	-0.16	-1.12	0.80
4	0	4401	15.6	-0.14	-0.94	0.76
8	0	703	2.5	-0.22	-2.20	1.14
9	0	1224	4.3	-0.16	-1.39	0.99
r	0	667	2.4	-0.14	-1.60	1.13

MM175

Response	Score	Count	Per cent	Pt Bis	Mean Ability	SD Ability
0	0	1057	3.7	-0.01	-0.62	0.92
1	0	1805	6.4	0.05	-0.40	0.95
2	0	10058	35.7	-0.18	-0.82	0.97
3	0	3495	12.4	0.20	-0.11	0.99
4	**1**	**5064**	**18.0**	**0.24**	**0.06**	**0.98**
8	0	3473	12.3	-0.07	-0.75	1.04
9	0	3135	11.1	-0.16	-1.04	0.95
r	0	118	0.4	-0.12	-2.77	1.00

MM191

Response	Score	Count	Per cent	Pt Bis	Mean Ability	SD Ability
1	0	3115	11.0	0.03	-0.51	0.93
2	0	12065	42.8	-0.02	-0.59	0.93
3	0	4061	14.4	-0.06	-0.70	0.87
4	**1**	**5346**	**19.0**	**0.28**	**0.11**	**1.04**
8	0	638	2.3	-0.20	-2.10	1.23
9	0	2828	10.0	-0.17	-1.12	1.04
r	0	149	0.5	-0.13	-2.67	1.06

MM197

Response	Score	Count	Per cent	Pt Bis	Mean Ability	SD Ability
1	0	4999	17.7	0.12	-0.34	0.92
2	0	11650	41.3	-0.18	-0.78	0.86
3	0	3773	13.4	0.09	-0.35	0.95
4	**1**	**5153**	**18.3**	**0.27**	**0.12**	**1.02**
8	0	1103	3.9	-0.19	-1.66	1.31
9	0	1283	4.6	-0.19	-1.51	1.05
r	0	228	0.8	-0.14	-2.36	1.06

MM206

Response	Score	Count	Per cent	Pt Bis	Mean Ability	SD Ability
0	0	910	3.2	0.04	-0.35	0.93
1	0	7519	26.7	0.06	-0.47	0.95
2	0	8827	31.3	0.14	-0.38	0.98
3	**1**	**1901**	**6.7**	**0.25**	**0.47**	**1.08**
8	0	4255	15.1	-0.18	-0.96	1.01
9	0	4548	16.1	-0.23	-1.08	0.93
r	0	246	0.9	-0.14	-2.30	1.07

MM208

Response	Score	Count	Per cent	Pt Bis	Mean Ability	SD Ability
1	0	4671	16.6	-0.15	-0.93	0.79
2	**1**	**15035**	**53.3**	**0.50**	**-0.05**	**0.91**
3	0	2104	7.5	-0.17	-1.17	0.79
4	0	4701	16.7	-0.22	-1.08	0.81
8	0	957	3.4	-0.24	-2.06	1.19
9	0	658	2.3	-0.13	-1.53	1.12
r	0	73	0.3	-0.10	-2.91	0.99

MM209

Response	Score	Count	Per cent	Pt Bis	Mean Ability	SD Ability
1	1	12672	45.0	0.16	-0.35	0.93
2	0	4668	16.6	-0.02	-0.71	0.91
3	0	2954	10.5	-0.05	-0.81	0.95
4	**2**	**4350**	**15.4**	**0.15**	**-0.02**	**0.99**
8	0	1165	4.1	-0.22	-1.82	1.10
9	0	2016	7.2	-0.20	-1.40	1.02
r	0	360	1.3	-0.14	-2.07	1.06

PM422

Response	Score	Count	Per cent	Pt Bis	Mean Ability	SD Ability
1	0	1391	4.9	-0.13	-1.14	0.82
2	**1**	**19739**	**70.0**	**0.49**	**-0.21**	**0.92**
3	0	2208	7.8	-0.17	-1.18	0.78
4	0	2254	8.0	-0.18	-1.21	0.74
5	0	1297	4.6	-0.18	-1.41	0.86
8	0	852	3.0	-0.26	-2.34	1.04
9	0	408	1.4	-0.13	-1.80	1.14
r	0	58	0.2	-0.09	-3.01	0.94

PM445

Response	Score	Count	Per cent	Pt Bis	Mean Ability	SD Ability
1	0	2568	9.1	-0.16	-1.08	0.73
2	0	4070	14.4	-0.16	-0.97	0.73
3	**1**	**13399**	**47.5**	**0.52**	**0.03**	**0.91**
4	0	3652	12.9	-0.19	-1.06	0.85
8	0	2706	9.6	-0.11	-1.03	1.09
9	0	1393	4.9	-0.19	-1.48	1.01
r	0	417	1.5	-0.15	-1.97	1.08

PM449

Response	Score	Count	Per cent	Pt Bis	Mean Ability	SD Ability
1	0	5532	19.6	-0.12	-0.81	0.83
2	0	6486	23.0	-0.11	-0.77	0.86
3	**1**	**8898**	**31.6**	**0.26**	**-0.09**	**0.99**
4	0	4707	16.7	0.16	-0.27	1.01
8	0	626	2.2	-0.20	-2.15	1.19
9	0	1769	6.3	-0.20	-1.41	1.05
r	0	183	0.6	-0.13	-2.52	1.05

PM454

Response	Score	Count	Per cent	Pt Bis	Mean Ability	SD Ability
1	**1**	**8715**	**30.9**	**0.45**	**0.17**	**0.97**
2	0	4789	17.0	-0.08	-0.77	0.77
3	0	3701	13.1	-0.13	-0.89	0.81
4	0	8673	30.8	-0.13	-0.78	0.88
8	0	654	2.3	-0.21	-2.20	1.22
9	0	1535	5.4	-0.13	-1.17	1.12
r	0	134	0.5	-0.12	-2.74	1.07

PM459

Response	Score	Count	Per cent	Pt Bis	Mean Ability	SD Ability
1	0	2284	8.1	-0.16	-1.11	0.73
2	0	2923	10.4	-0.13	-0.99	0.75
3	0	4999	17.7	-0.19	-1.00	0.80
4	**1**	**15666**	**55.6**	**0.50**	**-0.08**	**0.92**
8	0	684	2.4	-0.22	-2.20	1.17
9	0	1413	5.0	-0.20	-1.55	0.99
r	0	218	0.8	-0.14	-2.39	1.08

PM462

Response	Score	Count	Per cent	Pt Bis	Mean Ability	SD Ability
1	**1**	**18761**	**66.5**	**0.51**	**-0.17**	**0.91**
2	0	4049	14.4	-0.23	-1.15	0.78
3	0	2103	7.5	-0.19	-1.25	0.75
4	0	1739	6.2	-0.16	-1.21	0.72
8	0	554	2.0	-0.22	-2.47	1.09
9	0	856	3.0	-0.19	-1.75	1.03
r	0	139	0.5	-0.12	-2.70	1.07

PM468

Response	Score	Count	Per cent	Pt Bis	Mean Ability	SD Ability
1	0	2695	9.6	-0.16	-1.05	0.78
2	0	4222	15.0	-0.11	-0.85	0.77
3	**1**	**13986**	**49.6**	**0.49**	**-0.03**	**0.94**
4	0	5475	19.4	-0.22	-1.04	0.82
8	0	534	1.9	-0.21	-2.36	1.12
9	0	1204	4.3	-0.19	-1.54	1.05
r	0	91	0.3	-0.11	-2.86	0.99

PM469

Response	Score	Count	Per cent	Pt Bis	Mean Ability	SD Ability
1	**1**	**11261**	**39.9**	**0.50**	**0.09**	**0.93**
2	0	4409	15.6	-0.17	-0.95	0.80
3	0	4274	15.2	-0.05	-0.72	0.81
4	0	6772	24.0	-0.23	-0.98	0.80
8	0	549	1.9	-0.22	-2.46	1.11
9	0	790	2.8	-0.17	-1.67	1.06
r	0	145	0.5	-0.13	-2.68	1.07

PM942

Response	Score	Count	Per cent	Pt Bis	Mean Ability	SD Ability
1	0	3076	10.9	-0.02	-0.67	0.89
2	**1**	**12447**	**44.1**	**0.37**	**-0.08**	**0.97**
3	0	4707	16.7	-0.19	-1.01	0.86
4	0	5548	19.7	-0.05	-0.70	0.80
8	0	1068	3.8	-0.21	-1.82	1.20
9	0	1246	4.4	-0.19	-1.55	1.07
r	0	115	0.4	-0.12	-2.77	1.01

Appendix 1b – Item descriptive statistics – MILO Reading

MR001

Response	Score	Count	Per cent	Pt Bis	Mean Ability	SD Ability
1	0	4326	15.3	-0.29	-1.08	0.69
2	**1**	**19599**	**69.5**	**0.52**	**0.17**	**1.10**
3	0	1704	6.0	-0.18	-1.12	0.75
4	0	1322	4.7	-0.16	-1.10	0.73
8	0	933	3.3	-0.25	-2.14	1.25
9	0	244	0.9	-0.10	-1.72	1.23
r	0	78	0.3	-0.10	-3.16	0.74

MR002

Response	Score	Count	Per cent	Pt Bis	Mean Ability	SD Ability
1	0	1808	6.4	-0.18	-1.07	0.70
2	0	3761	13.3	-0.26	-1.06	0.70
3	0	2706	9.6	-0.21	-1.03	0.75
4	**1**	**18538**	**65.7**	**0.55**	**0.23**	**1.09**
8	0	918	3.3	-0.26	-2.20	1.24
9	0	387	1.4	-0.10	-1.31	1.13
r	0	87	0.3	-0.11	-3.19	0.81

MR003

Response	Score	Count	Per cent	Pt Bis	Mean Ability	SD Ability
1	0	2883	10.2	-0.20	-0.99	0.72
2	0	3202	11.4	-0.23	-1.02	0.74
3	**1**	**18154**	**64.4**	**0.53**	**0.23**	**1.11**
4	0	2812	10.0	-0.20	-0.99	0.74
8	0	793	2.8	-0.26	-2.38	1.17
9	0	273	1.0	-0.10	-1.60	1.15
r	0	88	0.3	-0.11	-3.18	0.81

MR024

Response	Score	Count	Per cent	Pt Bis	Mean Ability	SD Ability
1	**1**	**7350**	**26.1**	**0.50**	**0.78**	**1.25**
2	0	4907	17.4	-0.17	-0.66	0.89
3	0	10049	35.6	-0.09	-0.42	0.82
4	0	4208	14.9	-0.13	-0.61	0.91
8	0	898	3.2	-0.25	-2.14	1.31
9	0	668	2.4	-0.09	-1.01	1.33
r	0	124	0.4	-0.13	-2.96	0.89

MR025

Response	Score	Count	Per cent	Pt Bis	Mean Ability	SD Ability
1	0	5453	19.3	-0.14	-0.64	0.81
2	**1**	**14644**	**51.9**	**0.44**	**0.29**	**1.13**
3	0	3537	12.5	-0.21	-0.92	0.80
4	0	3084	10.9	-0.09	-0.60	1.09
8	0	828	2.9	-0.25	-2.31	1.23
9	0	526	1.9	-0.11	-1.30	1.17
r	0	133	0.5	-0.13	-2.91	0.90

MR033

Response	Score	Count	Per cent	Pt Bis	Mean Ability	SD Ability
1	**1**	**5647**	**31.2**	**0.22**	**-0.00**	**0.99**
2	0	3091	17.1	-0.10	-0.68	0.81
3	0	4754	26.3	0.03	-0.43	0.92
4	0	2583	14.3	0.07	-0.32	1.08
8	0	953	5.3	-0.26	-1.86	1.47
9	0	864	4.8	-0.15	-1.20	1.11
r	0	198	1.1	-0.15	-2.16	1.23

MR034

Response	Score	Count	Per cent	Pt Bis	Mean Ability	SD Ability
1	**1**	**2737**	**27.1**	**0.34**	**0.95**	**1.46**
2	0	2056	20.3	-0.19	-0.41	1.02
3	0	2475	24.5	-0.04	-0.05	1.20
4	0	2641	26.1	-0.10	-0.17	1.07
8	0	25	0.2	-0.06	-1.97	1.75
9	0	134	1.3	-0.06	-0.67	1.25
r	0	44	0.4	-0.12	-2.54	1.02

MR035

Response	Score	Count	Per cent	Pt Bis	Mean Ability	SD Ability
1	0	5146	18.2	0.01	-0.30	0.96
2	**1**	**9314**	**33.0**	**0.41**	**0.50**	**1.26**
3	0	7684	27.2	-0.20	-0.64	0.85
4	0	3700	13.1	-0.09	-0.55	0.86
8	0	861	3.1	-0.23	-2.02	1.39
9	0	1191	4.2	-0.11	-0.96	1.14
r	0	307	1.1	-0.13	-1.95	1.30

MR041

Response	Score	Count	Per cent	Pt Bis	Mean Ability	SD Ability
1	0	6675	23.7	0.05	-0.20	1.02
2	0	4734	16.8	-0.10	-0.53	0.94
3	0	5616	19.9	-0.14	-0.59	0.89
4	**1**	**8545**	**30.3**	**0.32**	**0.41**	**1.27**
8	0	943	3.3	-0.22	-1.92	1.42
9	0	1322	4.7	-0.11	-0.89	1.12
r	0	373	1.3	-0.13	-1.75	1.32

MR042

Response	Score	Count	Per cent	Pt Bis	Mean Ability	SD Ability
1	0	7443	26.4	-0.04	-0.37	0.91
2	0	5273	18.7	0.01	-0.27	0.99
3	0	5301	18.8	-0.13	-0.58	0.98
4	**1**	**7868**	**27.9**	**0.31**	**0.44**	**1.33**
8	0	898	3.2	-0.23	-2.00	1.39
9	0	1037	3.7	-0.12	-1.03	1.10
r	0	389	1.4	-0.13	-1.71	1.31

MR043

Response	Score	Count	Per cent	Pt Bis	Mean Ability	SD Ability
1	0	2930	10.4	-0.13	-0.78	0.95
2	**1**	**19007**	**67.4**	**0.41**	**0.11**	**1.13**
3	0	5396	19.1	-0.27	-0.93	0.89
8	0	471	1.7	-0.21	-2.64	1.28
9	0	306	1.1	-0.12	-1.76	1.18
r	0	91	0.3	-0.11	-3.14	0.83

MR044

Response	Score	Count	Per cent	Pt Bis	Mean Ability	SD Ability
1	0	7652	27.1	-0.05	-0.43	1.00
2	0	3586	12.7	-0.30	-1.18	0.78
3	**1**	**15934**	**56.5**	**0.34**	**0.15**	**1.15**
8	0	651	2.3	-0.20	-2.12	1.41
9	0	260	0.9	-0.11	-1.76	1.13
r	0	121	0.4	-0.12	-2.98	0.86

MR056

Response	Score	Count	Per cent	Pt Bis	Mean Ability	SD Ability
1	0	2725	9.7	-0.22	-1.07	0.68
2	0	2934	10.4	-0.24	-1.08	0.66

Response	Score	Count	Per cent	Pt Bis	Mean Ability	SD Ability
3	0	1764	6.3	-0.18	-1.08	0.76
4	**1**	**19541**	**69.3**	**0.53**	**0.18**	**1.09**
8	0	679	2.4	-0.26	-2.56	1.08
9	0	403	1.4	-0.13	-1.71	1.19
r	0	136	0.5	-0.13	-2.89	0.91

MR058

Response	Score	Count	Per cent	Pt Bis	Mean Ability	SD Ability
1	0	2702	9.6	-0.14	-0.76	0.78
2	0	2873	10.2	-0.20	-0.93	0.72
3	0	6762	24.0	-0.30	-0.88	0.71
4	**1**	**14301**	**50.7**	**0.60**	**0.45**	**1.10**
8	0	681	2.4	-0.25	-2.46	1.15
9	0	713	2.5	-0.13	-1.30	1.15
r	0	150	0.5	-0.13	-2.80	0.96

MR059

Response	Score	Count	Per cent	Pt Bis	Mean Ability	SD Ability
1	0	5714	20.3	-0.00	-0.34	0.93
2	**1**	**13053**	**46.3**	**0.38**	**0.29**	**1.20**
3	0	4354	15.4	-0.16	-0.71	0.84
4	0	3574	12.7	-0.19	-0.85	0.81
8	0	735	2.6	-0.26	-2.47	1.13
9	0	594	2.1	-0.12	-1.32	1.20
r	0	158	0.6	-0.13	-2.75	0.99

MR069

Response	Score	Count	Per cent	Pt Bis	Mean Ability	SD Ability
1	0	6265	22.2	-0.08	-0.49	0.94
2	**1**	**14892**	**52.8**	**0.39**	**0.23**	**1.15**
3	0	2243	8.0	-0.17	-0.94	0.81
4	0	2461	8.7	-0.13	-0.79	0.92

Response	Score	Count	Per cent	Pt Bis	Mean Ability	SD Ability
8	0	915	3.2	-0.24	-2.07	1.32
9	0	1028	3.6	-0.12	-1.05	1.11
r	0	400	1.4	-0.13	-1.69	1.31

MR087

Response	Score	Count	Per cent	Pt Bis	Mean Ability	SD Ability
1	**1**	**8980**	**31.8**	**0.46**	**0.59**	**1.29**
2	0	9491	33.7	-0.01	-0.33	0.85
3	0	4025	14.3	-0.23	-0.87	0.76
4	0	4129	14.6	-0.15	-0.69	0.75
8	0	849	3.0	-0.25	-2.19	1.27
9	0	525	1.9	-0.14	-1.51	1.18
r	0	196	0.7	-0.14	-2.52	1.10

MR089

Response	Score	Count	Per cent	Pt Bis	Mean Ability	SD Ability
1	0	5850	20.7	-0.08	-0.46	0.84
2	0	5283	18.7	-0.18	-0.71	0.79
3	0	6171	21.9	-0.13	-0.56	0.88
4	**1**	**9175**	**32.5**	**0.50**	**0.64**	**1.21**
8	0	789	2.8	-0.25	-2.30	1.22
9	0	712	2.5	-0.14	-1.34	1.13
r	0	215	0.8	-0.14	-2.39	1.14

MR090

Response	Score	Count	Per cent	Pt Bis	Mean Ability	SD Ability
1	0	5479	19.4	-0.12	-0.57	0.84
2	0	5129	18.2	-0.10	-0.55	0.87
3	**1**	**11280**	**40.0**	**0.44**	**0.44**	**1.23**
4	0	4542	16.1	-0.14	-0.65	0.80
8	0	787	2.8	-0.26	-2.34	1.21
9	0	749	2.7	-0.14	-1.31	1.09

Response	Score	Count	Per cent	Pt Bis	Mean Ability	SD Ability
r	0	229	0.8	-0.14	-2.31	1.18

MR201

Response	Score	Count	Per cent	Pt Bis	Mean Ability	SD Ability
1	0	2683	9.5	-0.12	-0.69	0.87
2	0	7676	27.2	-0.11	-0.51	0.90
3	**1**	**12225**	**43.3**	**0.38**	**0.32**	**1.22**
4	0	3732	13.2	-0.12	-0.65	0.87
8	0	763	2.7	-0.24	-2.24	1.35
9	0	967	3.4	-0.10	-0.99	1.20
r	0	162	0.6	-0.13	-2.72	0.99

MR202

Response	Score	Count	Per cent	Pt Bis	Mean Ability	SD Ability
1	0	5904	20.9	-0.04	-0.38	0.85
2	0	5958	21.1	-0.20	-0.71	0.80
3	0	4615	16.4	-0.07	-0.48	0.91
4	**1**	**9741**	**34.5**	**0.41**	**0.48**	**1.29**
8	0	886	3.1	-0.23	-2.04	1.36
9	0	927	3.3	-0.12	-1.09	1.18
r	0	177	0.6	-0.14	-2.63	1.04

MR203

Response	Score	Count	Per cent	Pt Bis	Mean Ability	SD Ability
1	**1**	**13174**	**46.7**	**0.31**	**0.21**	**1.20**
2	0	3468	12.3	-0.06	-0.50	0.96
3	0	4873	17.3	-0.02	-0.39	0.99
4	0	4085	14.5	-0.17	-0.76	0.92
8	0	705	2.5	-0.23	-2.31	1.36
9	0	1721	6.1	-0.08	-0.71	1.08
r	0	182	0.6	-0.13	-2.59	1.06

MR204

Response	Score	Count	Per cent	Pt Bis	Mean Ability	SD Ability
1	0	3098	11.0	-0.08	-0.53	0.95
2	**1**	**8606**	**30.5**	**0.19**	**0.20**	**1.32**
3	0	10427	37.0	0.04	-0.25	1.01
4	0	4090	14.5	-0.02	-0.36	1.05
8	0	756	2.7	-0.25	-2.33	1.22
9	0	1040	3.7	-0.12	-1.06	1.17
r	0	190	0.7	-0.14	-2.53	1.07

PF449

Response	Score	Count	Per cent	Pt Bis	Mean Ability	SD Ability
1	0	1429	5.1	-0.16	-1.09	0.75
2	0	1985	7.0	-0.09	-0.76	0.96
3	**1**	**20412**	**72.4**	**0.37**	**0.05**	**1.15**
4	0	961	3.4	-0.12	-1.07	0.76
8	0	658	2.3	-0.17	-1.88	1.38
9	0	2712	9.6	-0.18	-1.03	1.07
r	0	49	0.2	-0.08	-3.16	0.69

PF455

Response	Score	Count	Per cent	Pt Bis	Mean Ability	SD Ability
1	0	4802	17.0	-0.12	-0.58	0.87
2	0	5288	18.7	-0.19	-0.70	0.80
3	0	5058	17.9	-0.13	-0.60	0.86
4	**1**	**10443**	**37.0**	**0.51**	**0.57**	**1.17**
8	0	944	3.3	-0.22	-1.87	1.46
9	0	1134	4.0	-0.12	-1.01	1.10
r	0	537	1.9	-0.12	-1.41	1.34

PF456

Response	Score	Count	Per cent	Pt Bis	Mean Ability	SD Ability
1	0	7497	26.6	-0.03	-0.36	0.92
2	0	4862	17.2	-0.09	-0.50	0.91
3	**1**	**8947**	**31.7**	**0.31**	**0.38**	**1.33**
4	0	4171	14.8	-0.06	-0.46	0.96
8	0	917	3.3	-0.20	-1.85	1.49
9	0	1213	4.3	-0.12	-0.98	1.11
r	0	599	2.1	-0.11	-1.31	1.34

PF458

Response	Score	Count	Per cent	Pt Bis	Mean Ability	SD Ability
1	**1**	**3884**	**21.5**	**0.28**	**0.20**	**1.11**
2	0	3461	19.1	0.05	-0.39	0.92
3	0	4205	23.2	-0.02	-0.49	0.92
4	0	2679	14.8	-0.05	-0.58	0.84
8	0	719	4.0	-0.28	-2.14	1.27
9	0	2590	14.3	-0.08	-0.70	1.03
r	0	554	3.1	-0.11	-1.24	1.30

PF487

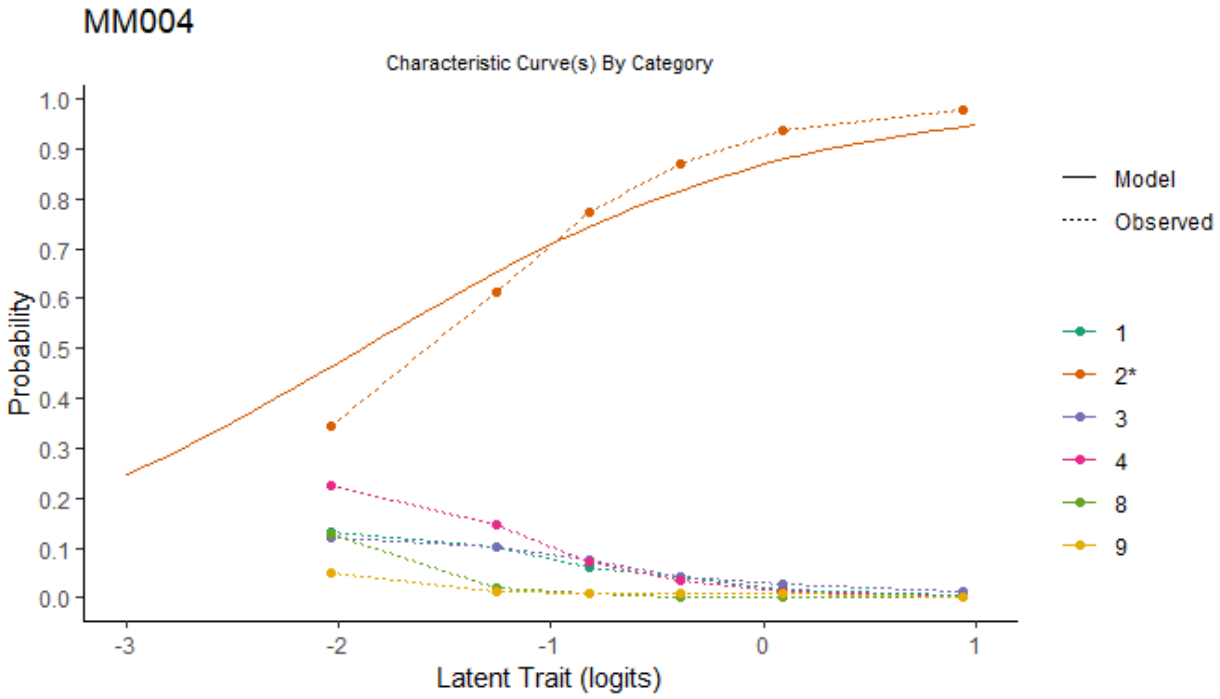
Response	Score	Count	Per cent	Pt Bis	Mean Ability	SD Ability
1	0	1411	5.0	-0.16	-1.08	0.73
2	**1**	**15933**	**56.5**	**0.35**	**0.16**	**1.18**
3	0	8215	29.1	-0.12	-0.55	0.95
4	0	1559	5.5	-0.14	-0.97	0.73
8	0	791	2.8	-0.24	-2.26	1.31
9	0	222	0.8	-0.08	-1.42	1.21
r	0	74	0.3	-0.10	-3.19	0.73

PF489

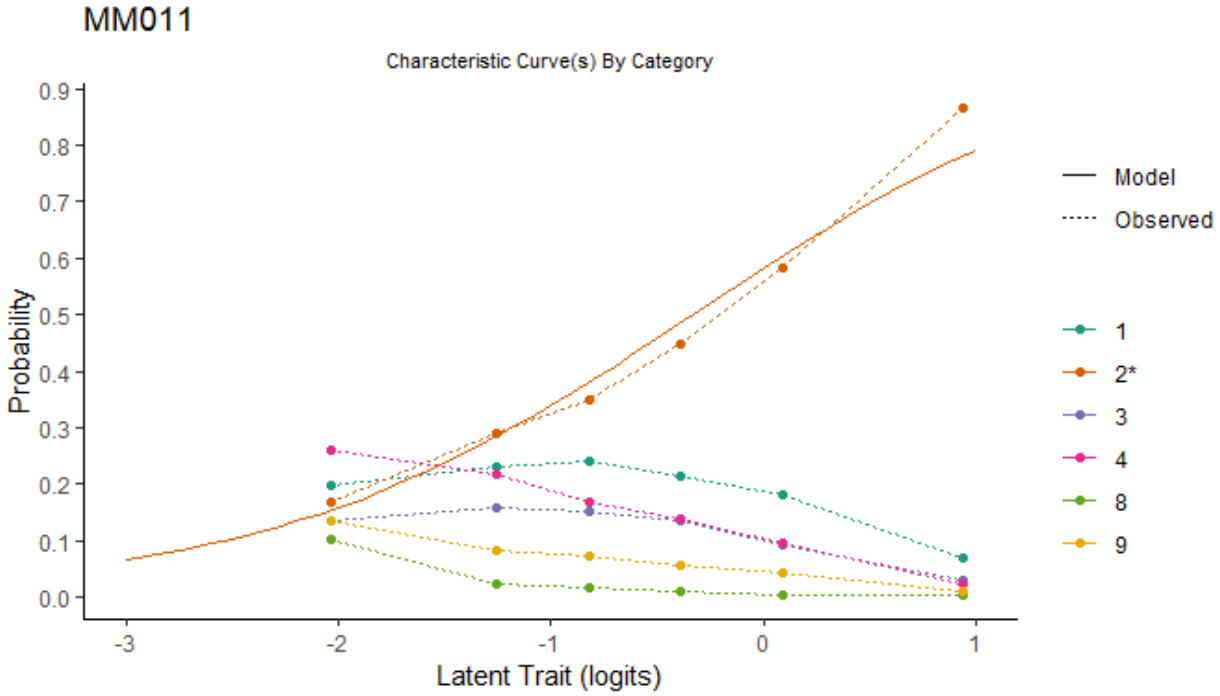
Response	Score	Count	Per cent	Pt Bis	Mean Ability	SD Ability
1	0	2936	10.4	-0.16	-0.80	1.00
2	**1**	**10235**	**36.3**	**0.36**	**0.39**	**1.28**
3	0	9741	34.5	-0.04	-0.38	0.91
4	0	3917	13.9	-0.14	-0.68	0.93
8	0	1041	3.7	-0.24	-1.95	1.34
9	0	259	0.9	-0.08	-1.28	1.19
r	0	76	0.3	-0.11	-3.19	0.73

Appendix 2a – Item Characteristic Curve (ICC) Graphs – MILO Mathematics

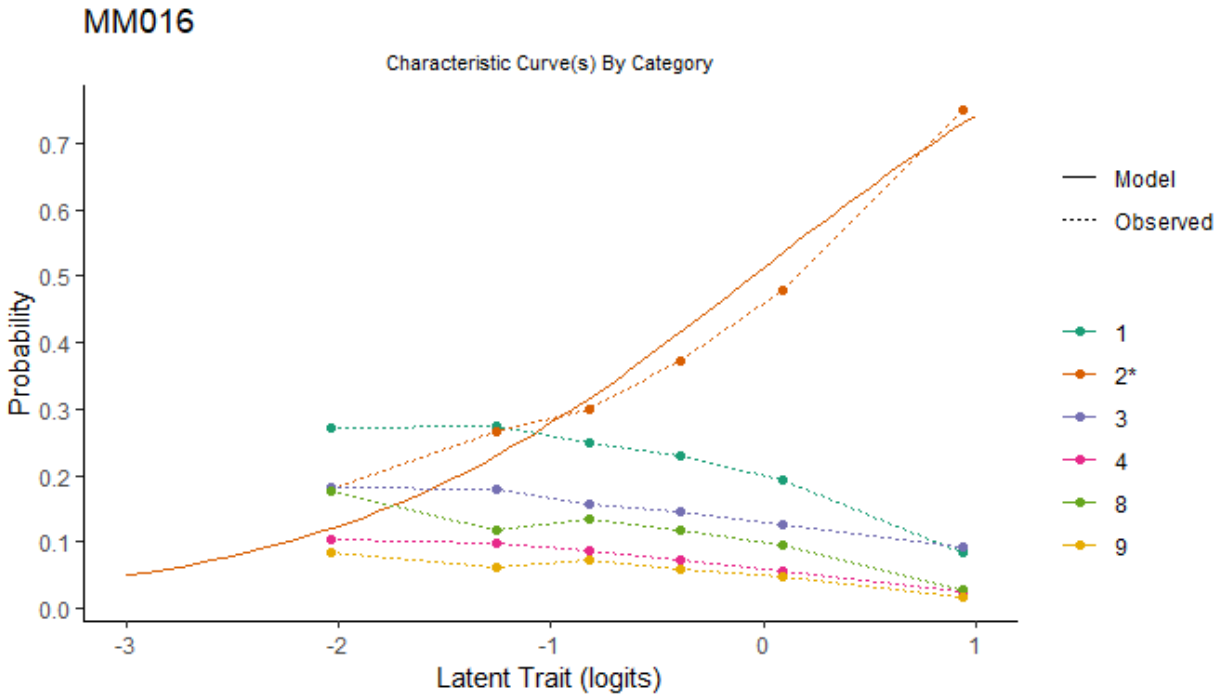
MM004



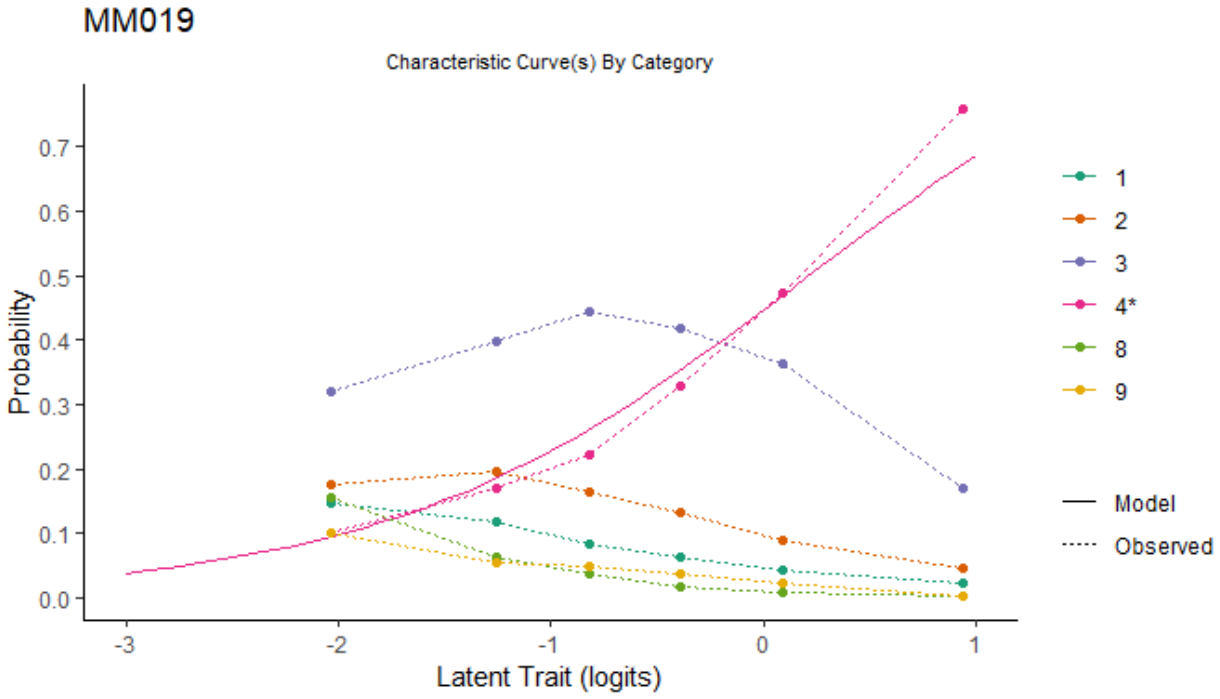
MM011



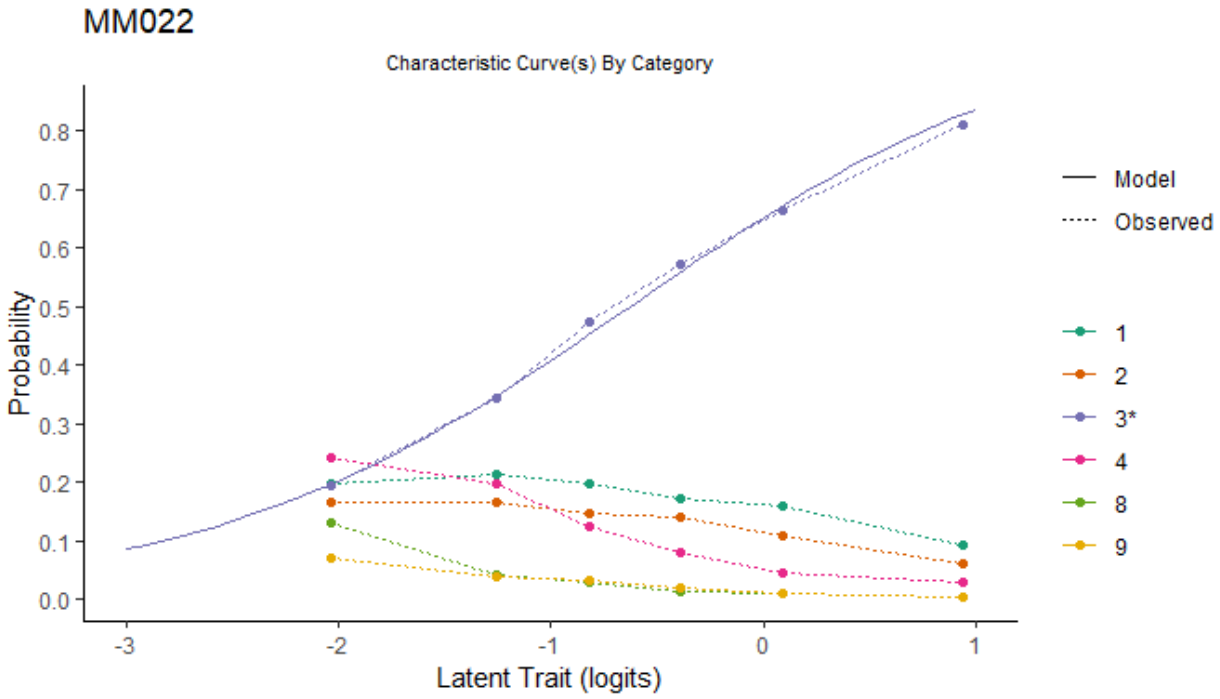
MM016


MM016
Characteristic Curve(s) By Category
Probability
Latent Trait (logits)
Model
Observed
1
2*
3
4
8
9

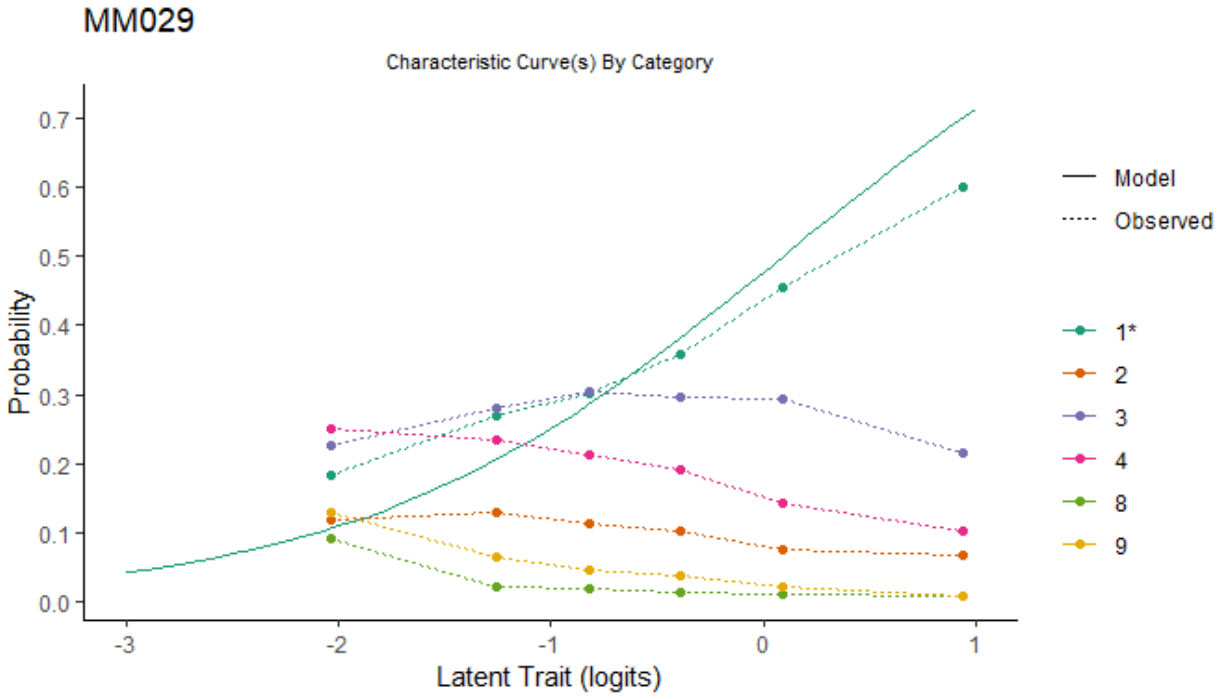
MM019


MM019
Characteristic Curve(s) By Category
Probability
Latent Trait (logits)
1
2
3
4*
8
9
Model
Observed

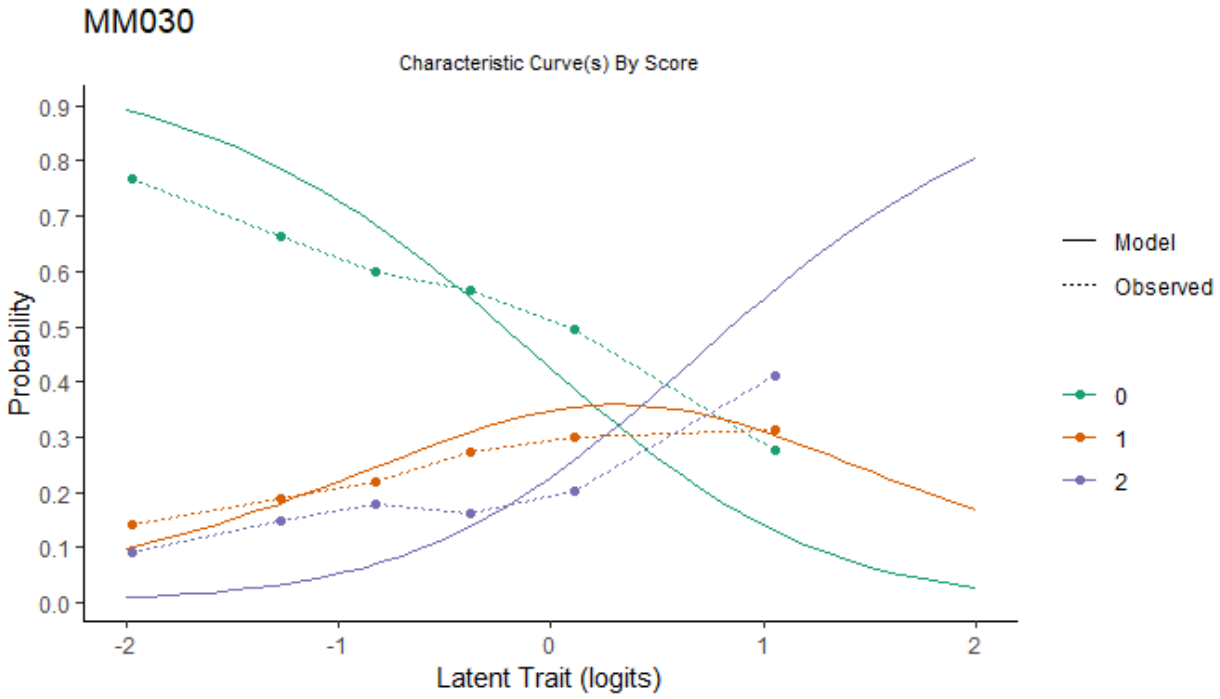
MM022


MM022
Characteristic Curve(s) By Category
Probability
Latent Trait (logits)
Model
Observed
1
2
3*
4
8
9

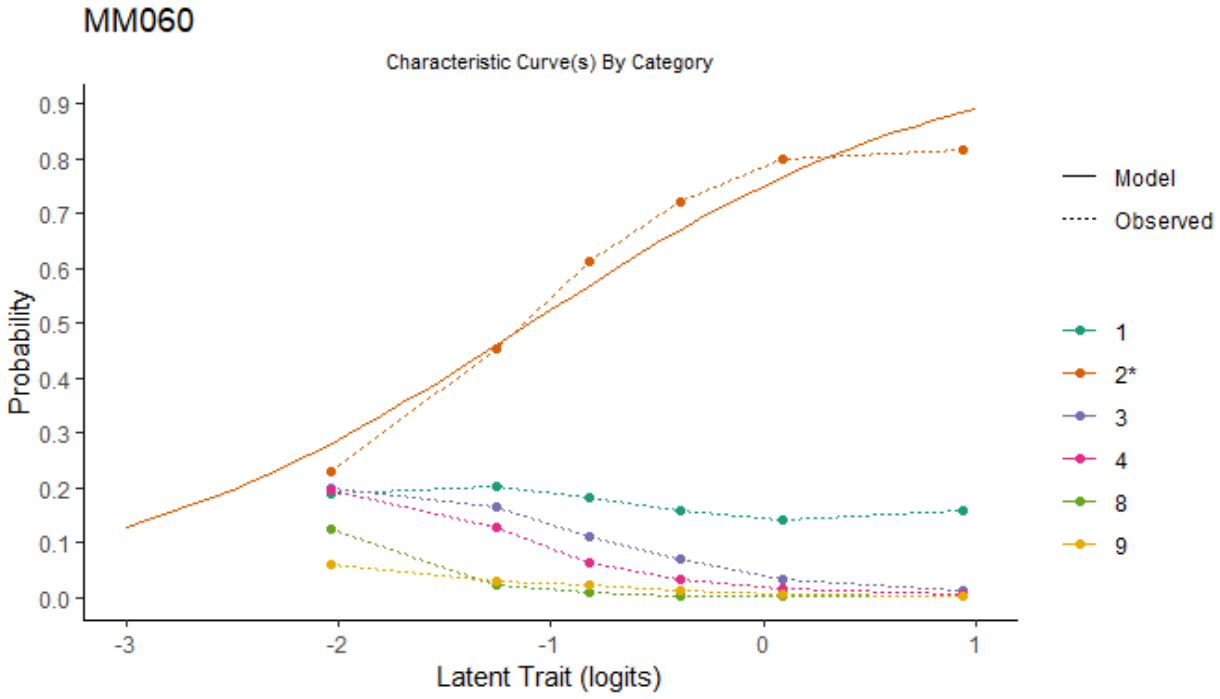
MM029


MM029
Characteristic Curve(s) By Category
Probability
Latent Trait (logits)
Model
Observed
1*
2
3
4
8
9

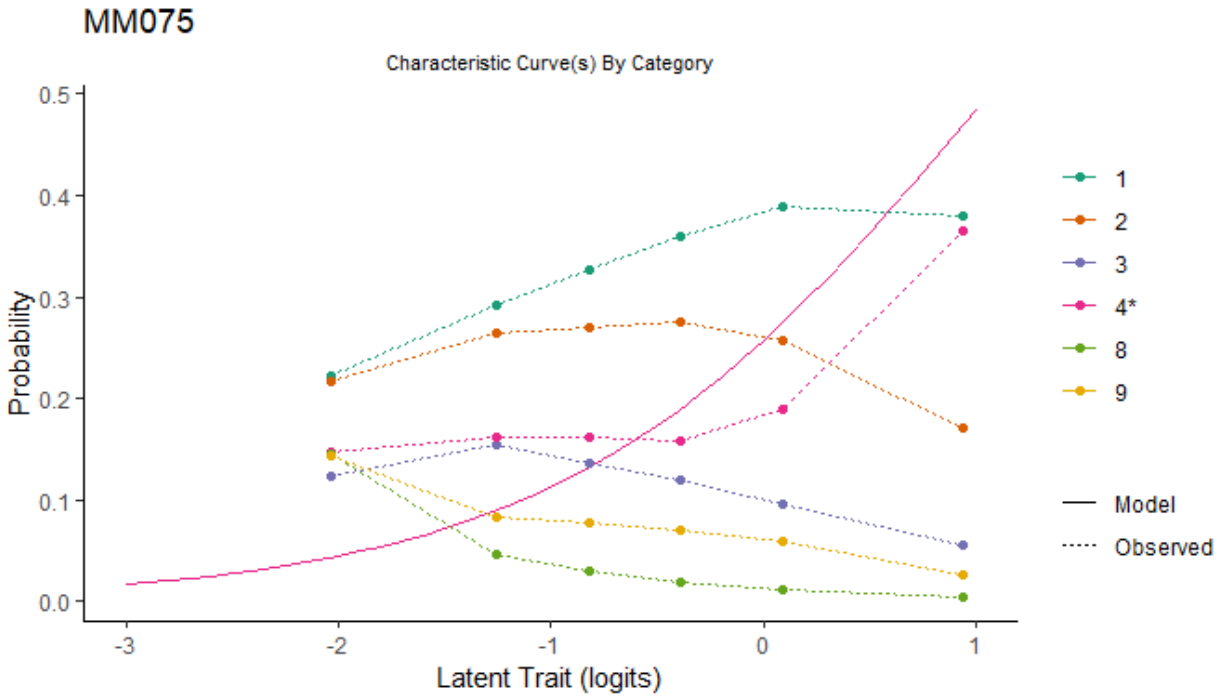
MM030


MM030
Characteristic Curve(s) By Score
Probability
Latent Trait (logits)
Model
Observed
0
1
2

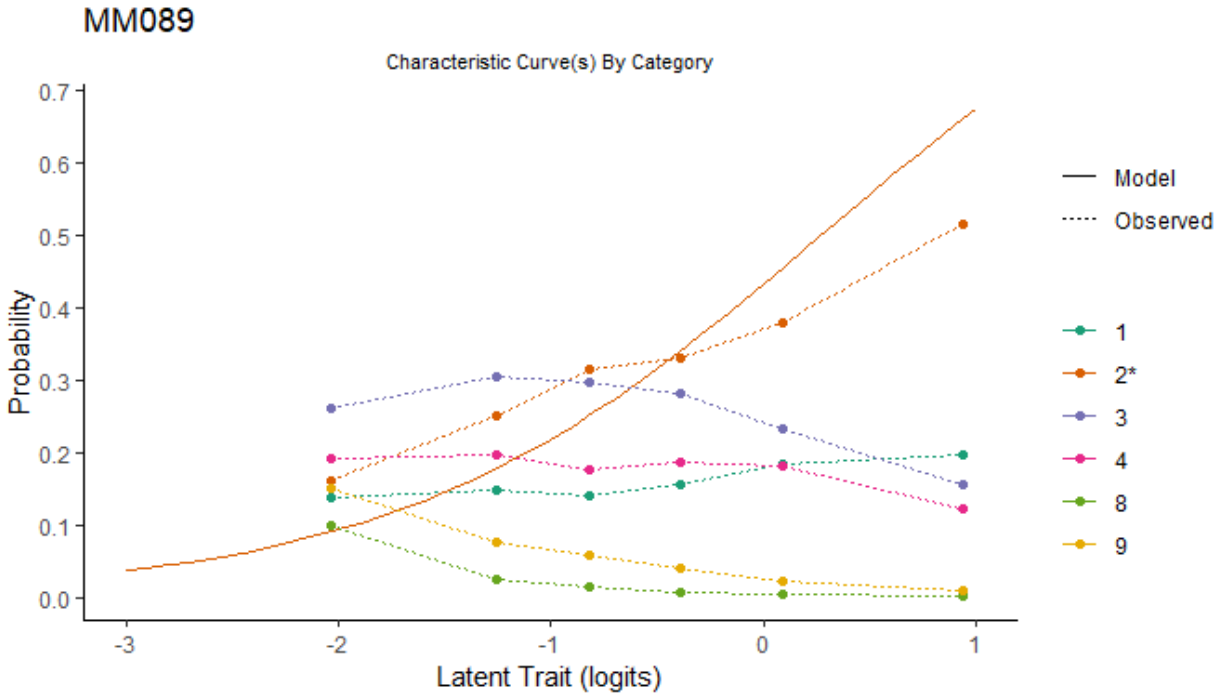
MM060


MM060
Characteristic Curve(s) By Category
Probability
Latent Trait (logits)
Model
Observed
1
2*
3
4
8
9

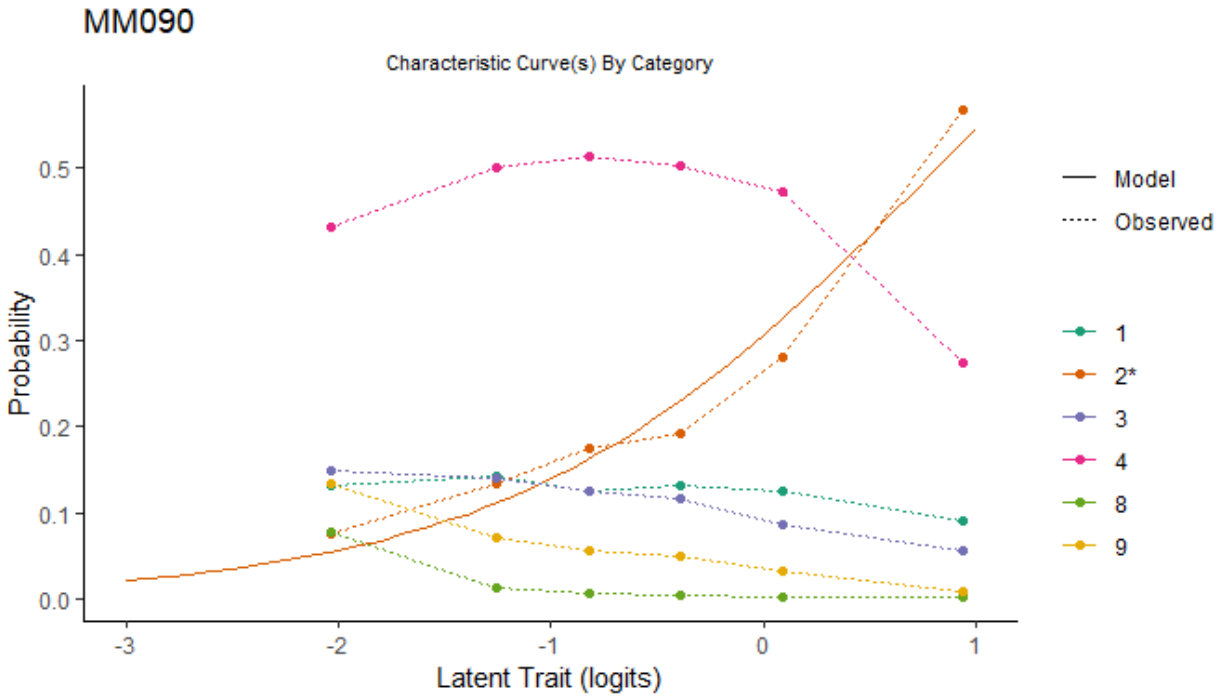
MM075


MM075
Characteristic Curve(s) By Category
Probability
0.5
0.4
0.3
0.2
0.1
0.0
-3
-2
-1
0
1
Latent Trait (logits)
1
2
3
4*
8
9
Model
Observed

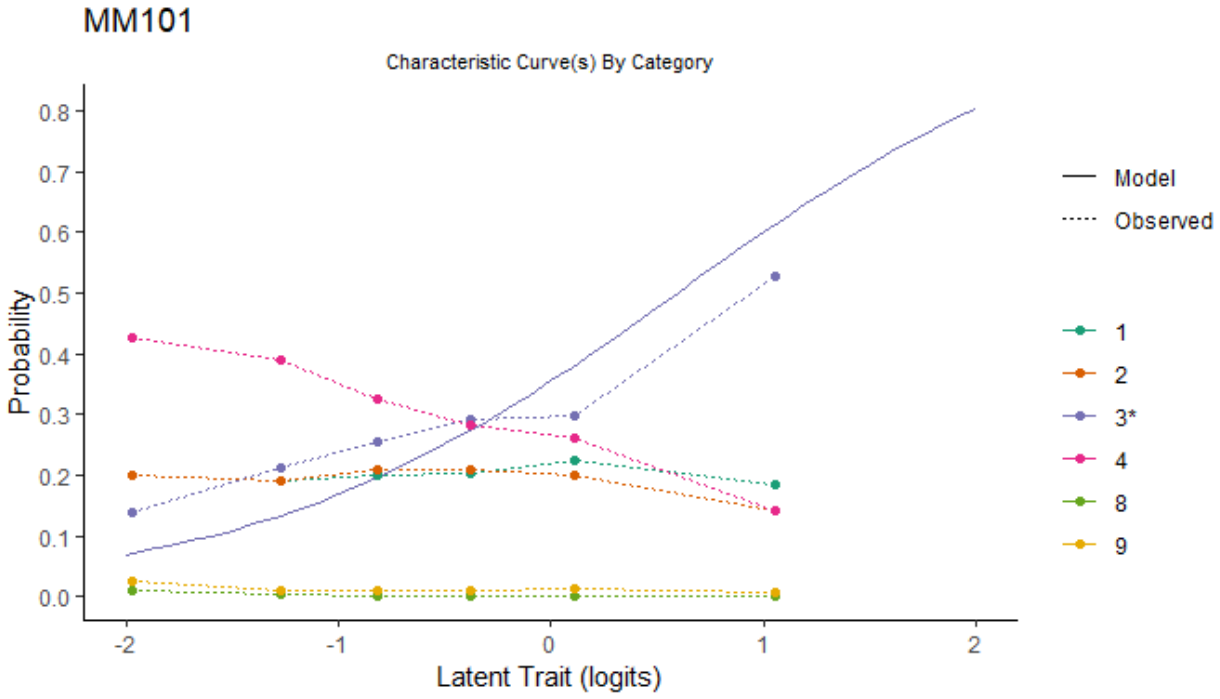
MM089


MM089
Characteristic Curve(s) By Category
Probability
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
-3
-2
-1
0
1
Latent Trait (logits)
Model
Observed
1
2*
3
4
8
9

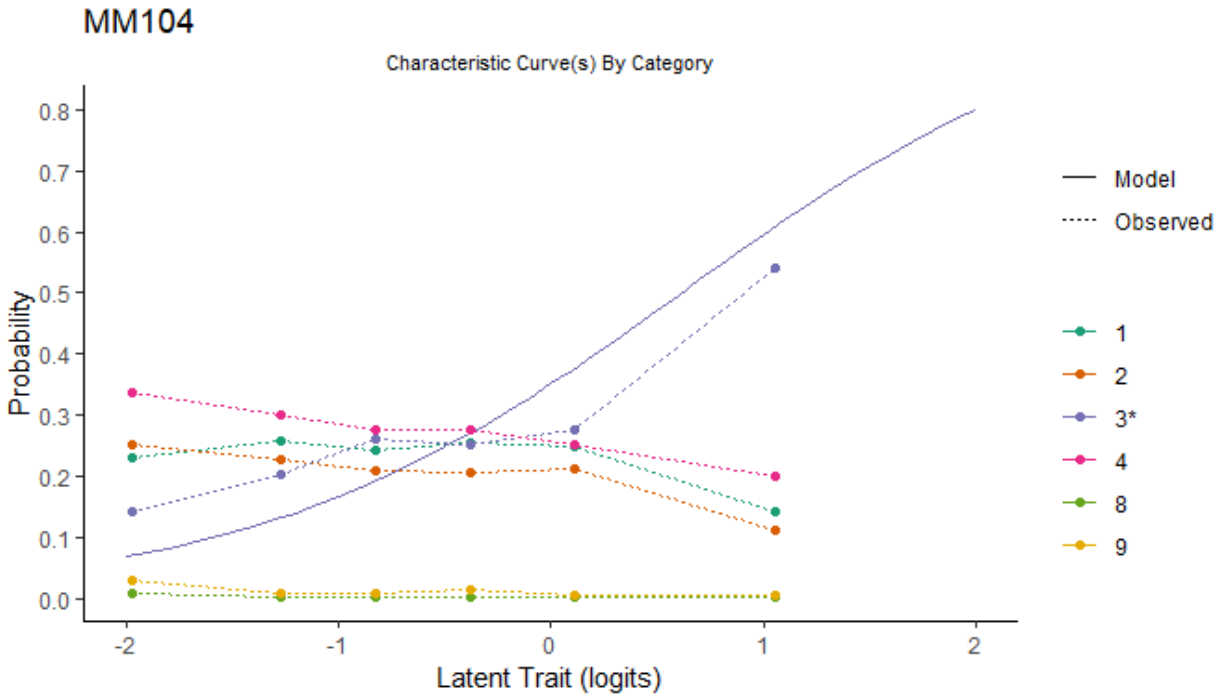
MM090


MM090
Characteristic Curve(s) By Category
Probability
Latent Trait (logits)
Model
Observed
1
2*
3
4
8
9

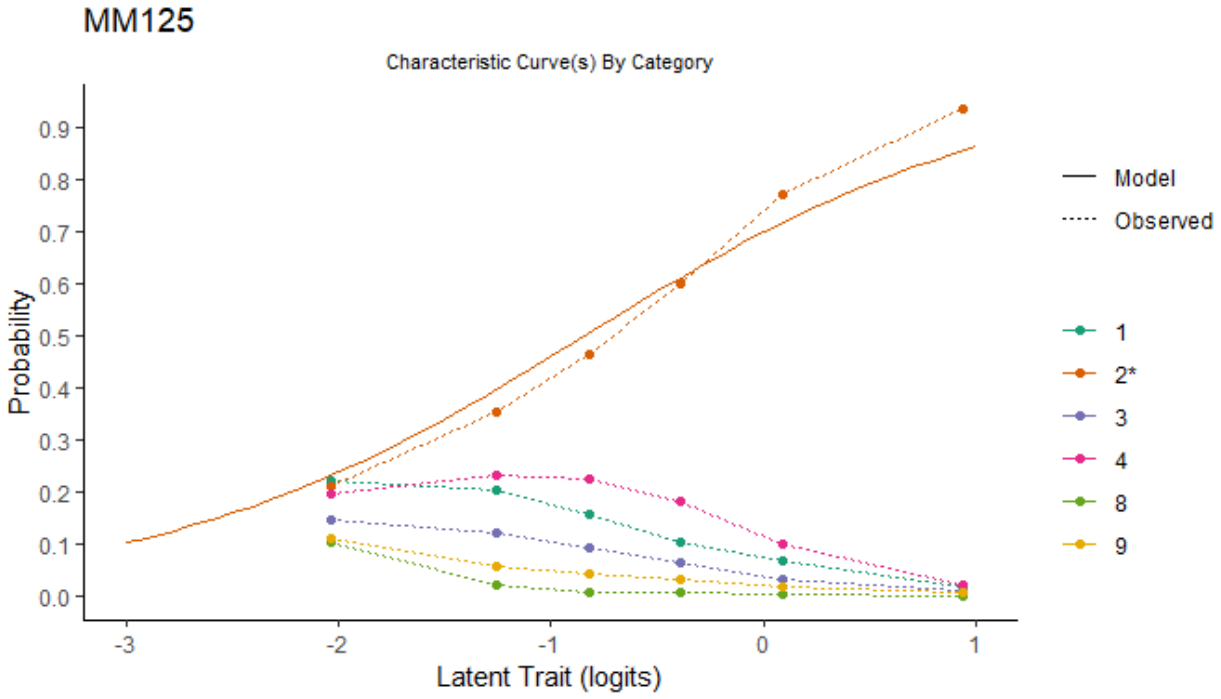
MM101


MM101
Characteristic Curve(s) By Category
Probability
Latent Trait (logits)
Model
Observed
1
2
3*
4
8
9

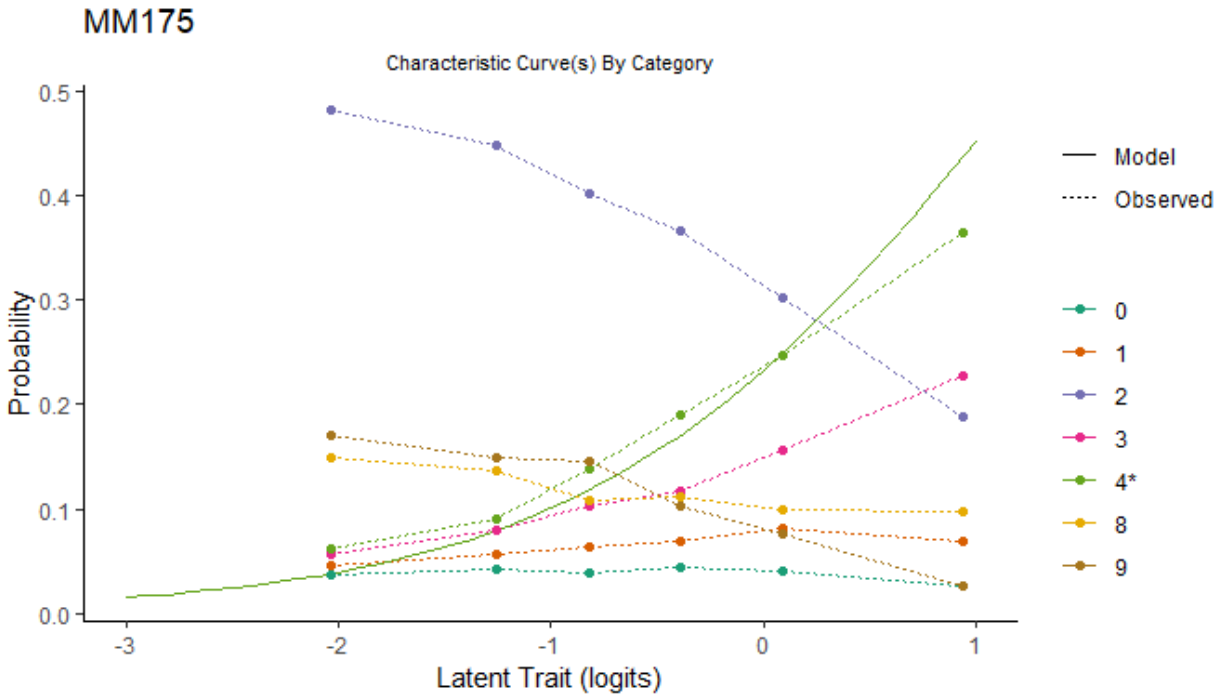
MM104


MM104
Characteristic Curve(s) By Category
Probability
Latent Trait (logits)
Model
Observed
1
2
3*
4
8
9

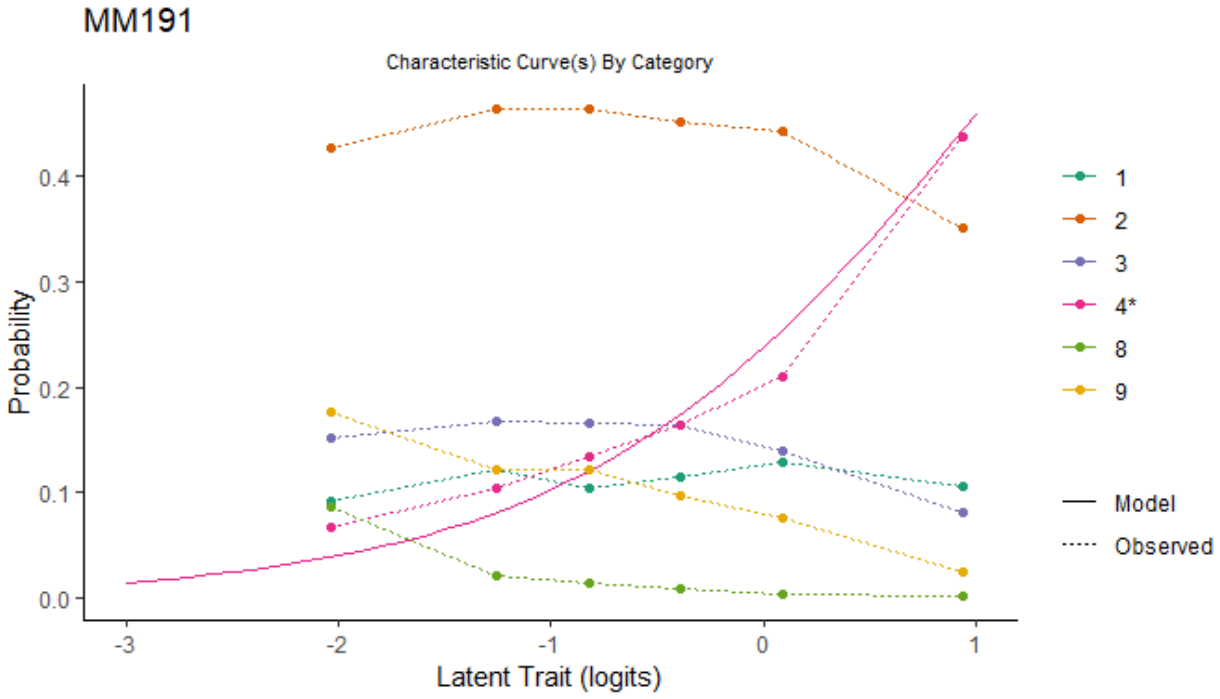
MM125


MM125
Characteristic Curve(s) By Category
Probability
Latent Trait (logits)
Model
Observed
1
2*
3
4
8
9

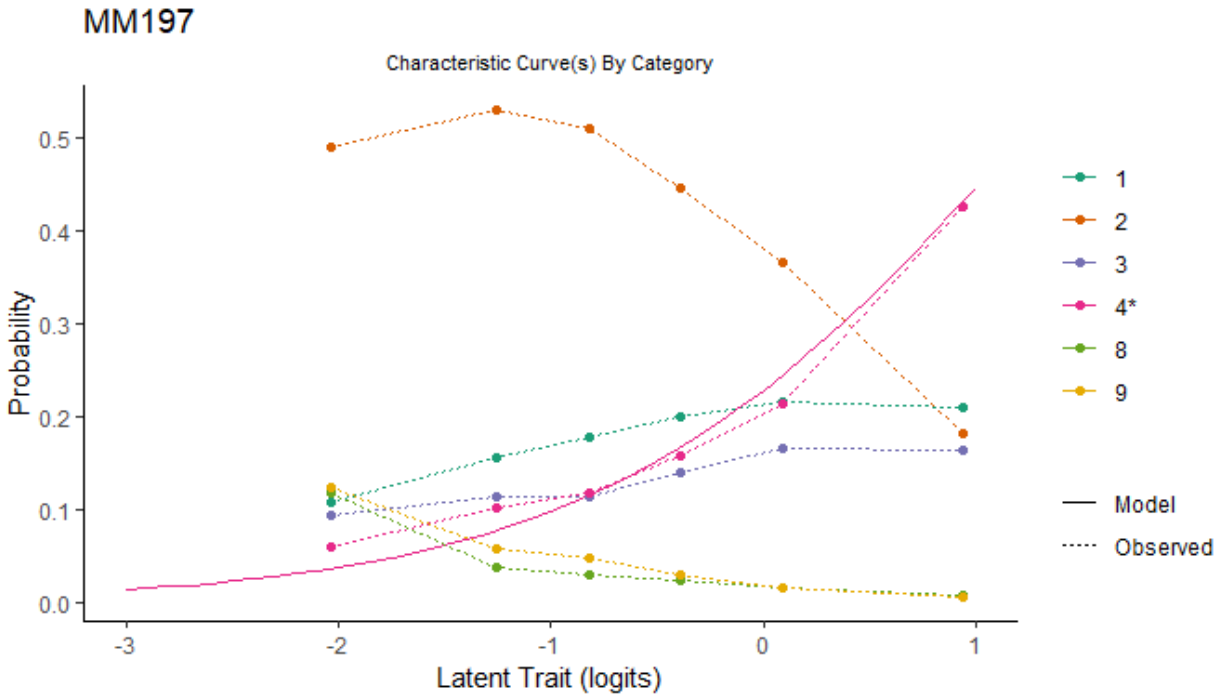
MM175


MM175
Characteristic Curve(s) By Category
Probability
0.0
0.1
0.2
0.3
0.4
0.5
-3
-2
-1
0
1
Latent Trait (logits)
Model
Observed
0
1
2
3
4*
8
9

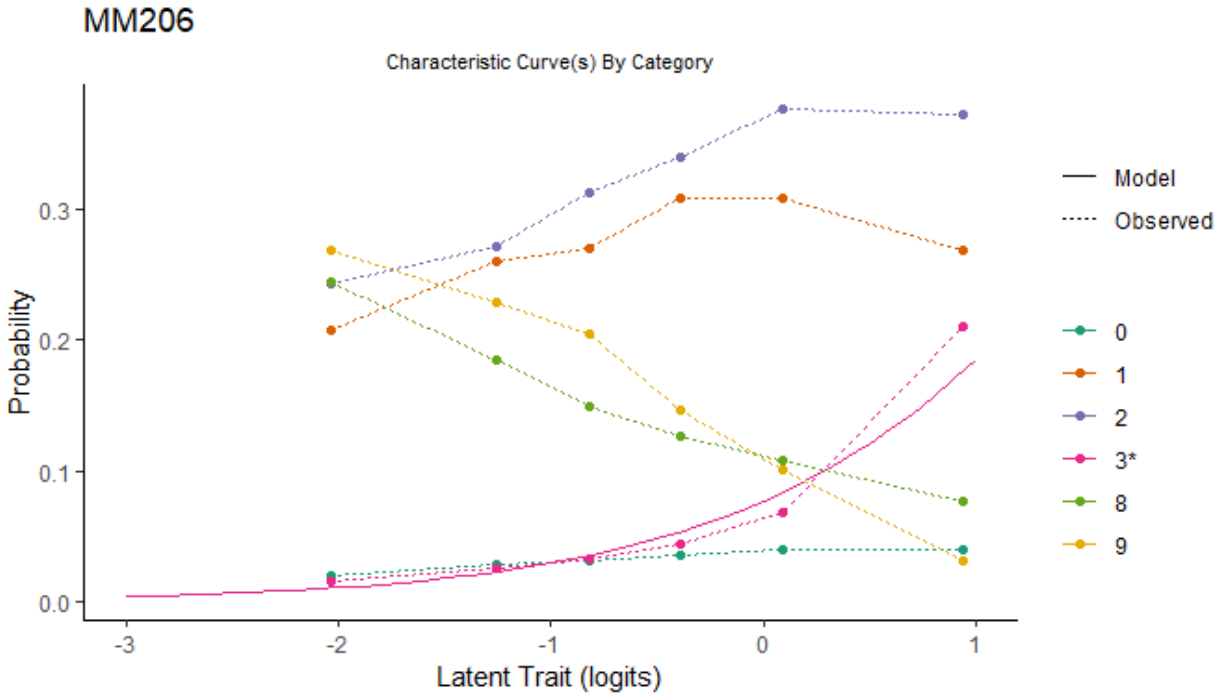
MM191


MM191
Characteristic Curve(s) By Category
Probability
0.0
0.1
0.2
0.3
0.4
-3
-2
-1
0
1
Latent Trait (logits)
1
2
3
4*
8
9
Model
Observed

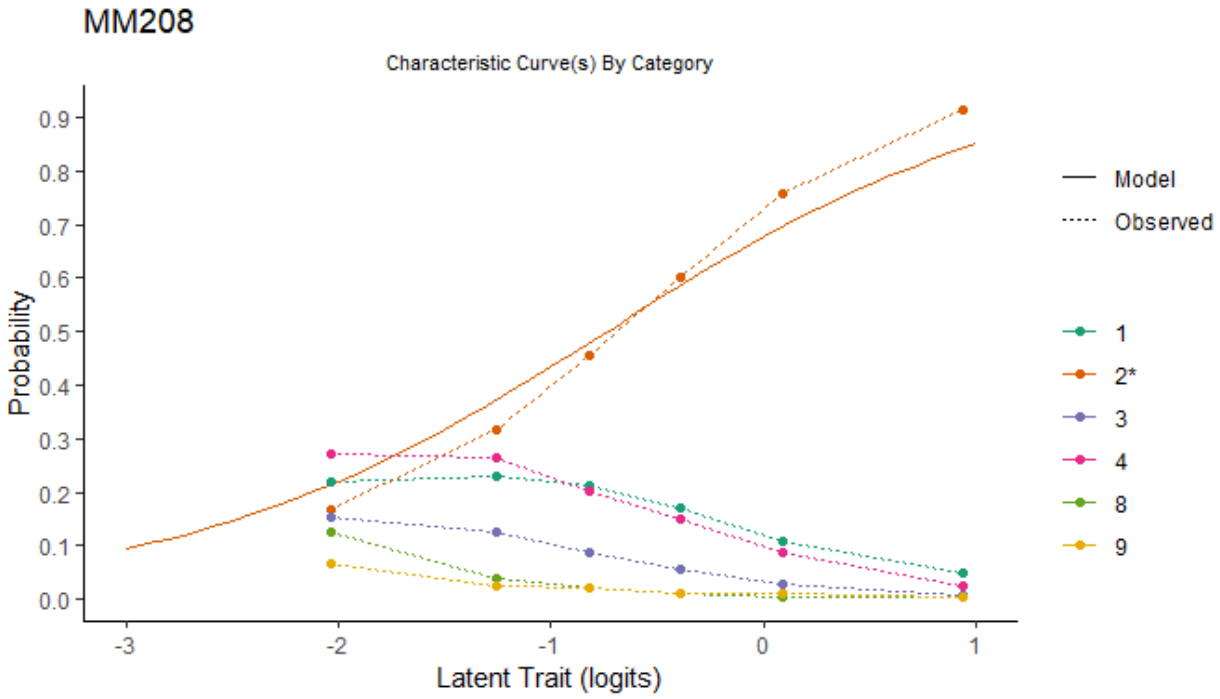
MM197


MM197
Characteristic Curve(s) By Category
Probability
0.0
0.1
0.2
0.3
0.4
0.5
-3
-2
-1
0
1
Latent Trait (logits)
1
2
3
4*
8
9
Model
Observed

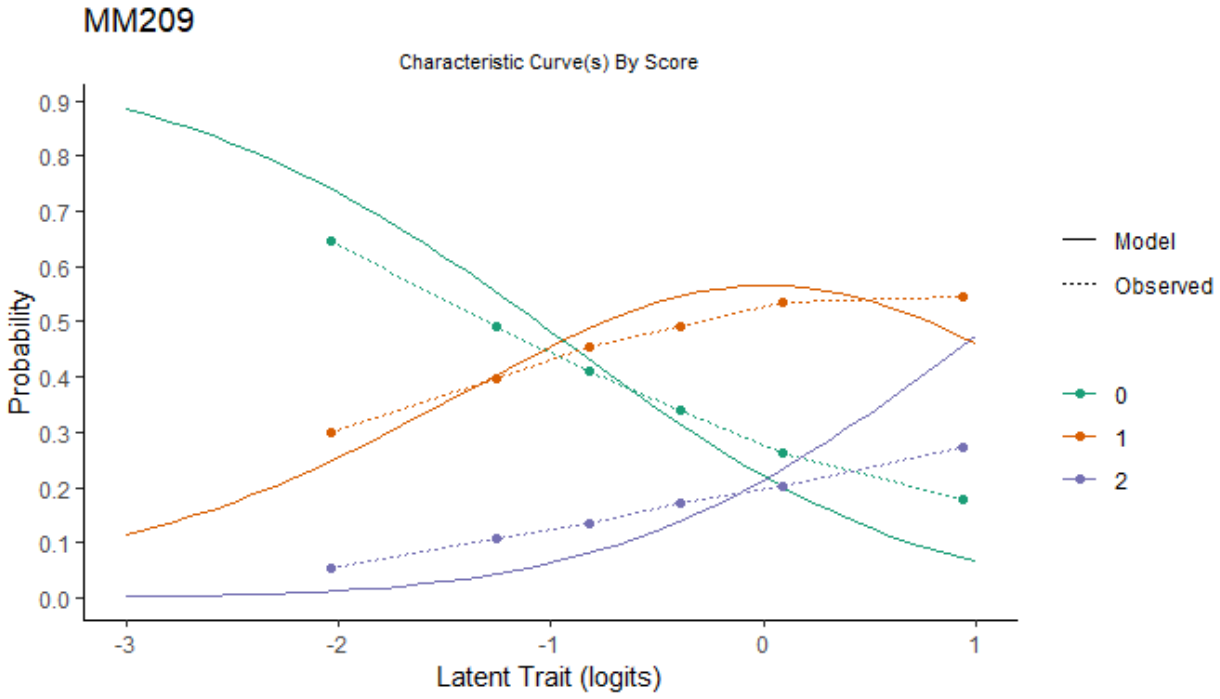
MM206


MM206
Characteristic Curve(s) By Category
Probability
0.0
0.1
0.2
0.3
-3
-2
-1
0
1
Latent Trait (logits)
Model
Observed
0
1
2
3*
8
9

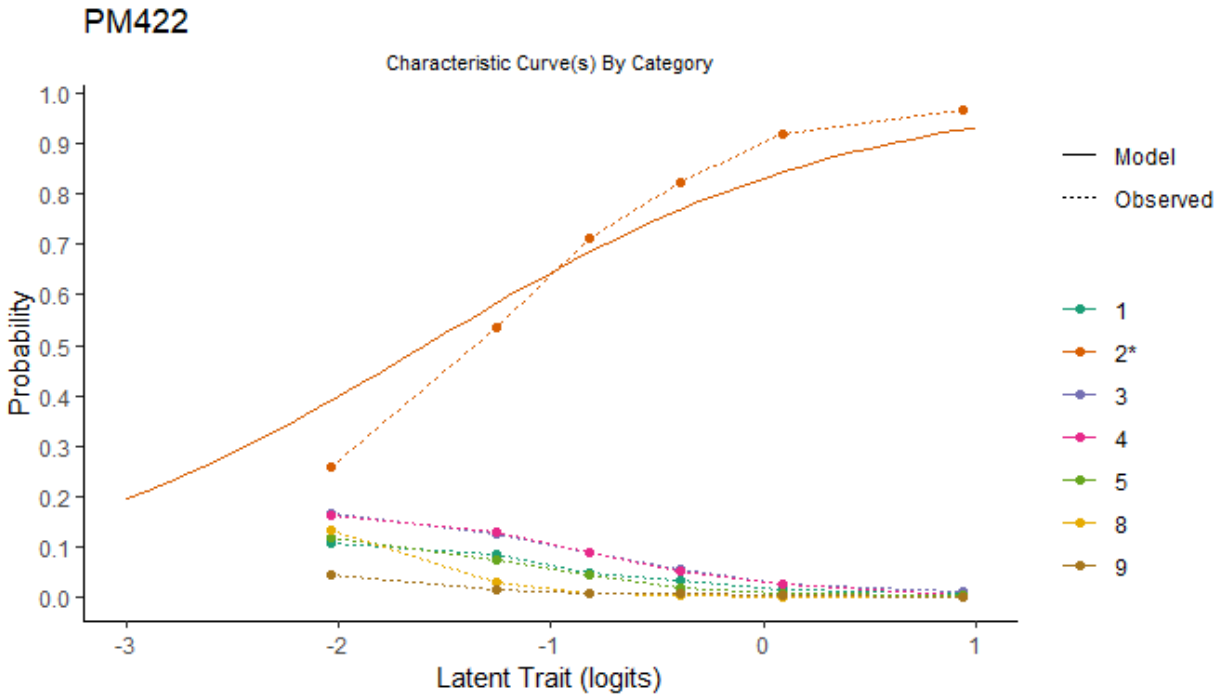
MM208


MM208
Characteristic Curve(s) By Category
Probability
Latent Trait (logits)
Model
Observed
1
2*
3
4
8
9

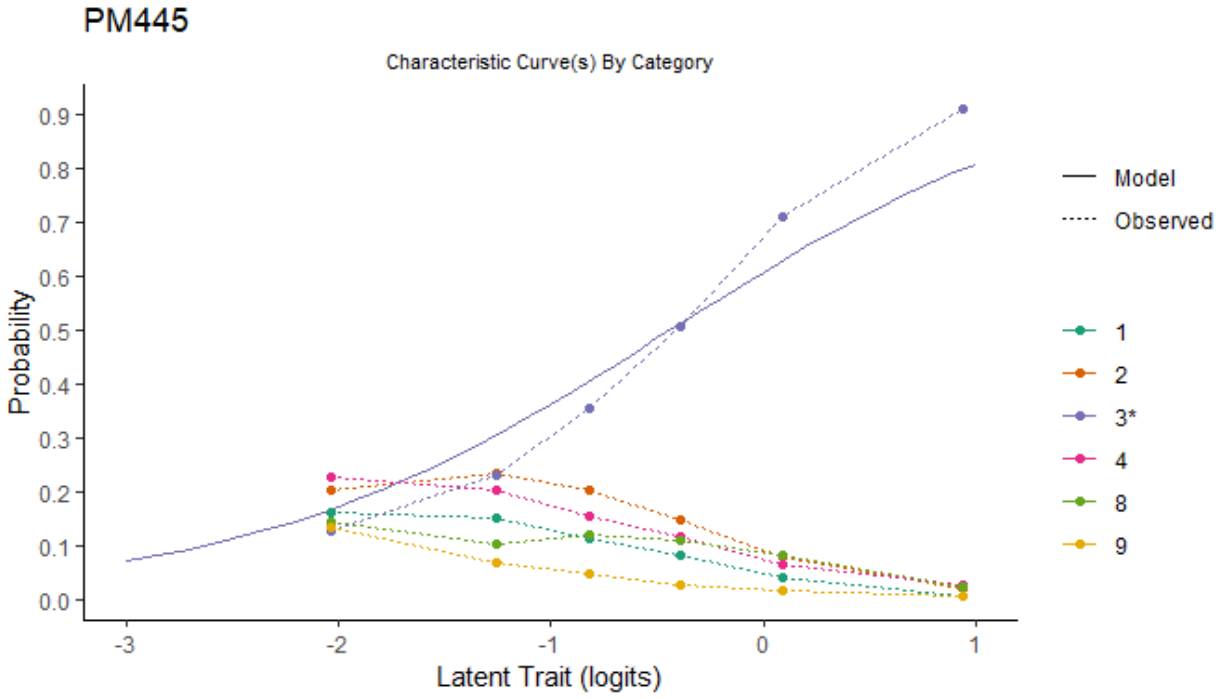
MM209


MM209
Characteristic Curve(s) By Score
Probability
Latent Trait (logits)
Model
Observed
0
1
2

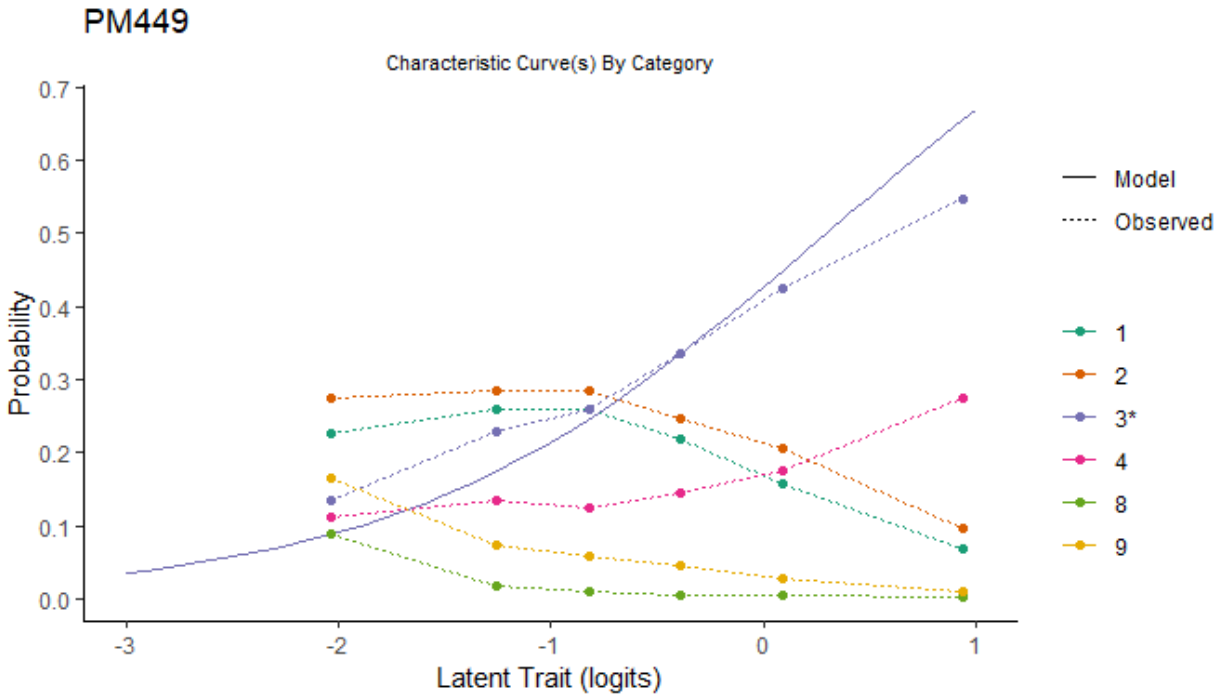
PM422


PM422
Characteristic Curve(s) By Category
Probability
Latent Trait (logits)
Model
Observed
1
2*
3
4
5
8
9

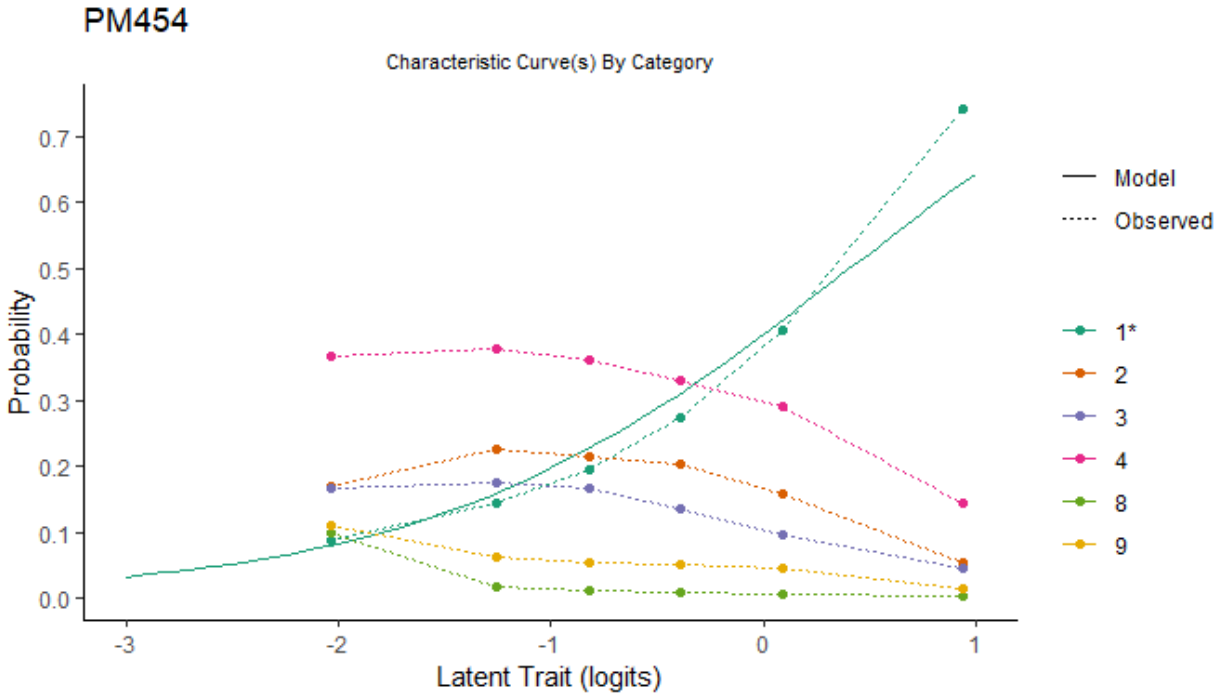
PM445


PM445
Characteristic Curve(s) By Category
Probability
Latent Trait (logits)
Model
Observed
1
2
3*
4
8
9

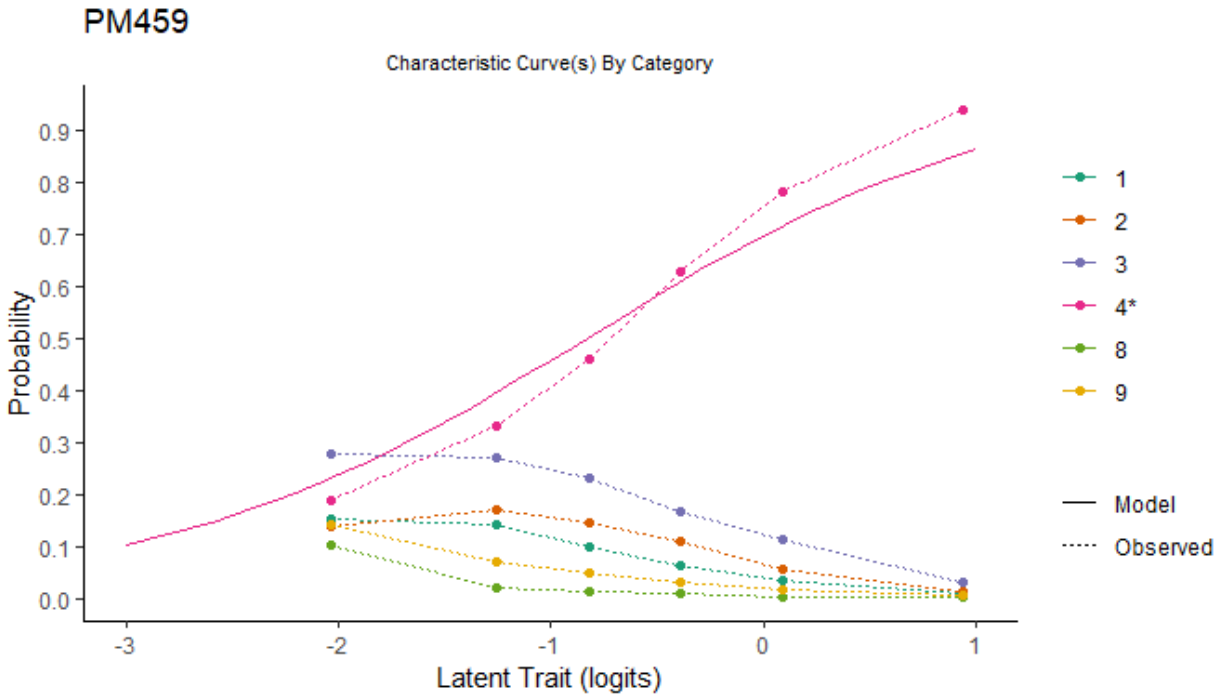
PM449


PM449
Characteristic Curve(s) By Category
Probability
Latent Trait (logits)
Model
Observed
1
2
3*
4
8
9

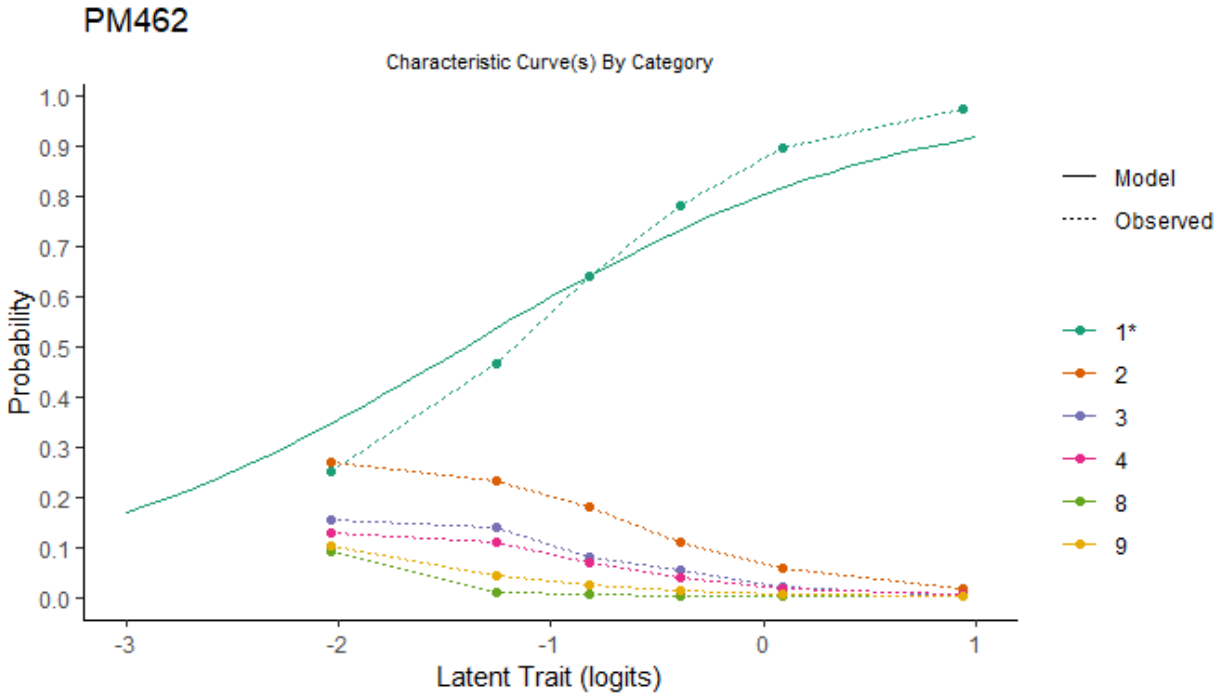
PM454


PM454
Characteristic Curve(s) By Category
Probability
Latent Trait (logits)
Model
Observed
1*
2
3
4
8
9

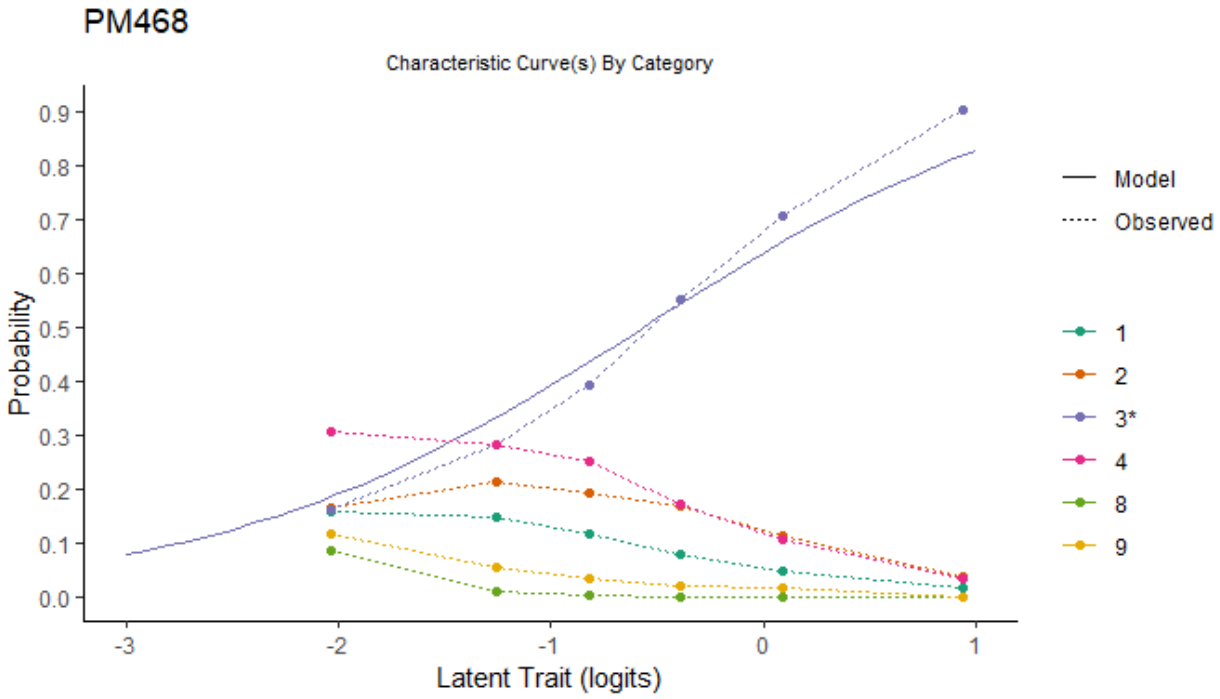
PM459


PM459
Characteristic Curve(s) By Category
Probability
Latent Trait (logits)
1
2
3
4*
8
9
Model
Observed

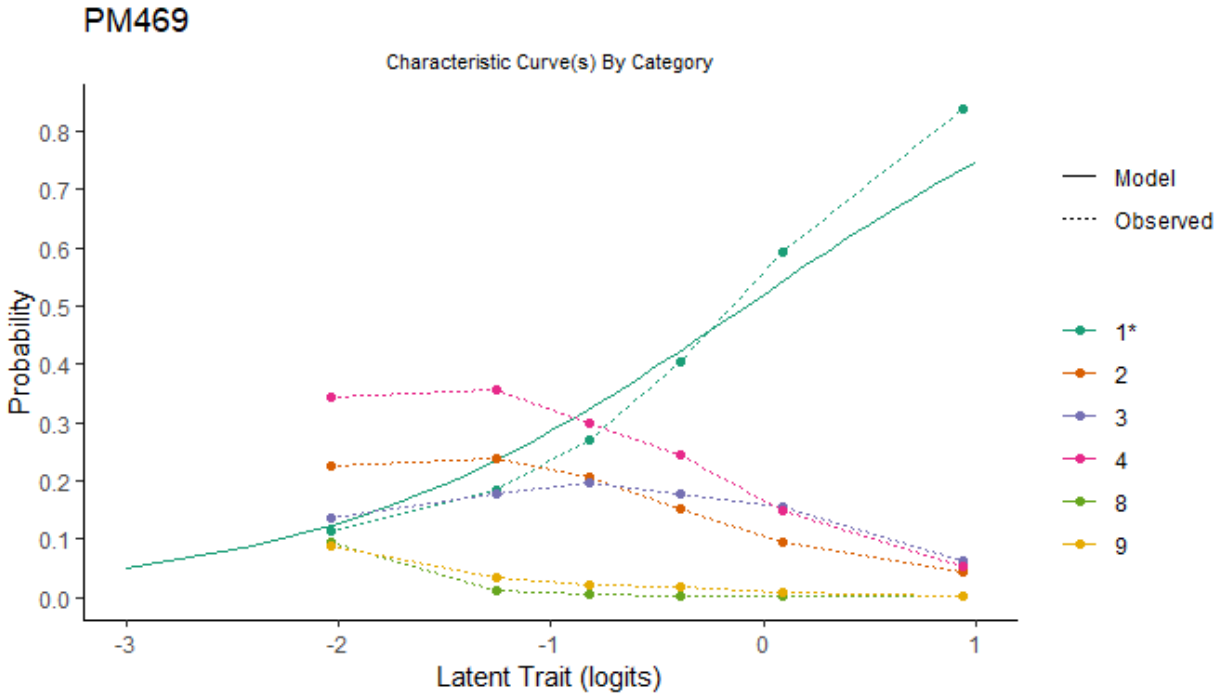
PM462


PM462
Characteristic Curve(s) By Category
Probability
Latent Trait (logits)
Model
Observed
1*
2
3
4
8
9

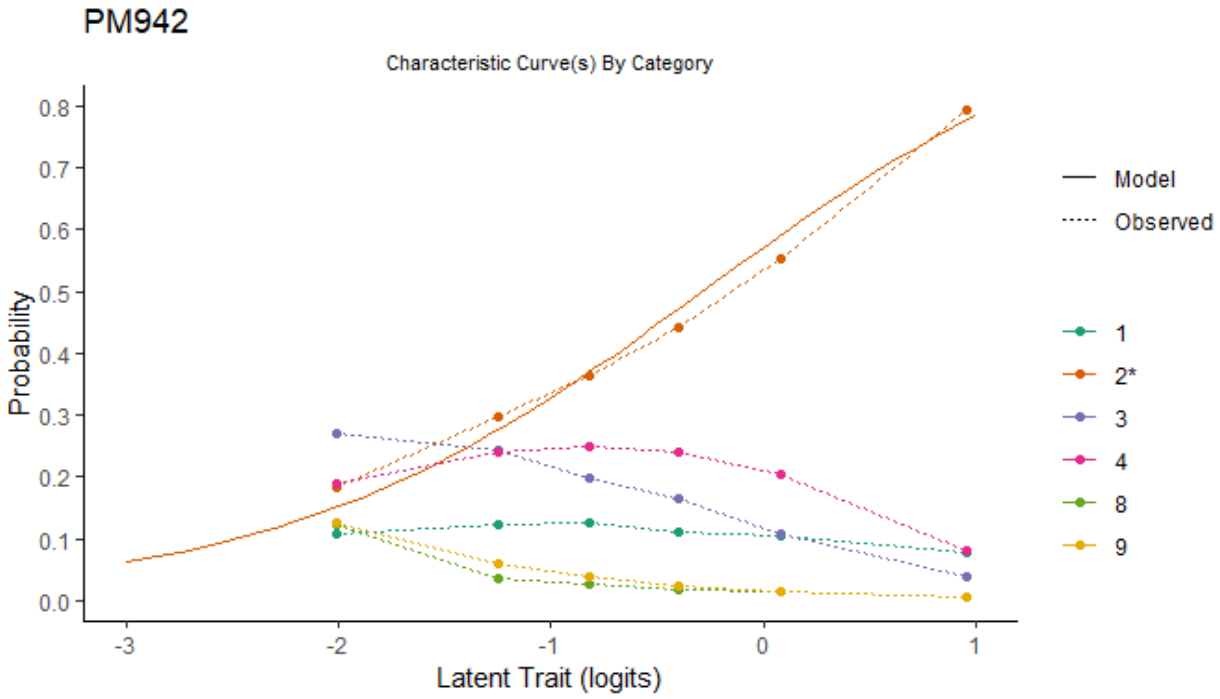
PM468


PM468
Characteristic Curve(s) By Category
Probability
Latent Trait (logits)
Model
Observed
1
2
3*
4
8
9

PM469

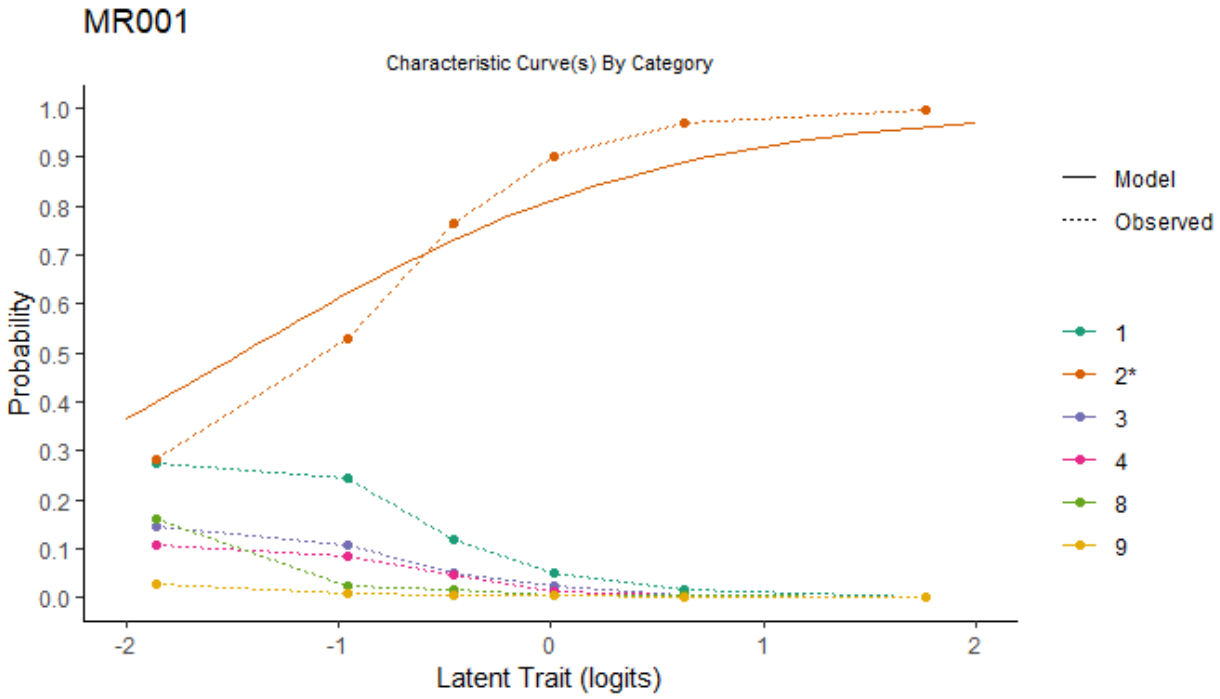

PM469
Characteristic Curve(s) By Category
Probability
Latent Trait (logits)
Model
Observed
1*
2
3
4
8
9

PM942

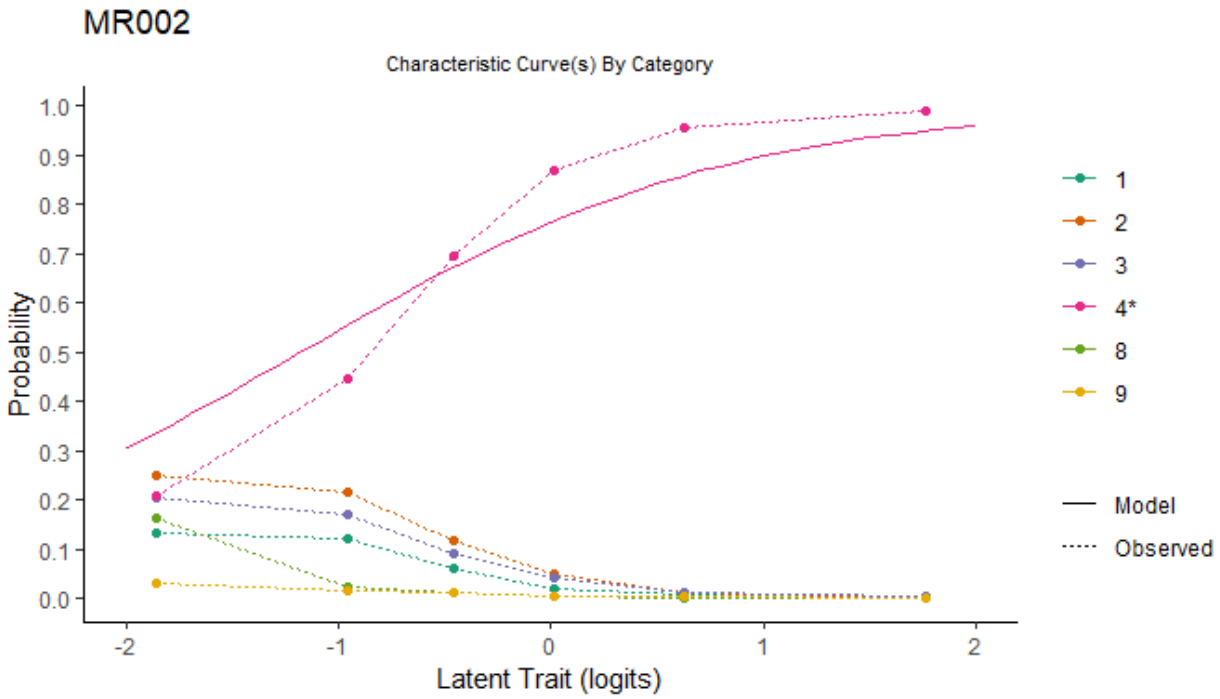

PM942
Characteristic Curve(s) By Category
Probability
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
-3
-2
-1
0
1
Latent Trait (logits)
Model
Observed
1
2*
3
4
8
9

Appendix 2b – Item Characteristic Curve (ICC) Graphs – MILO Reading

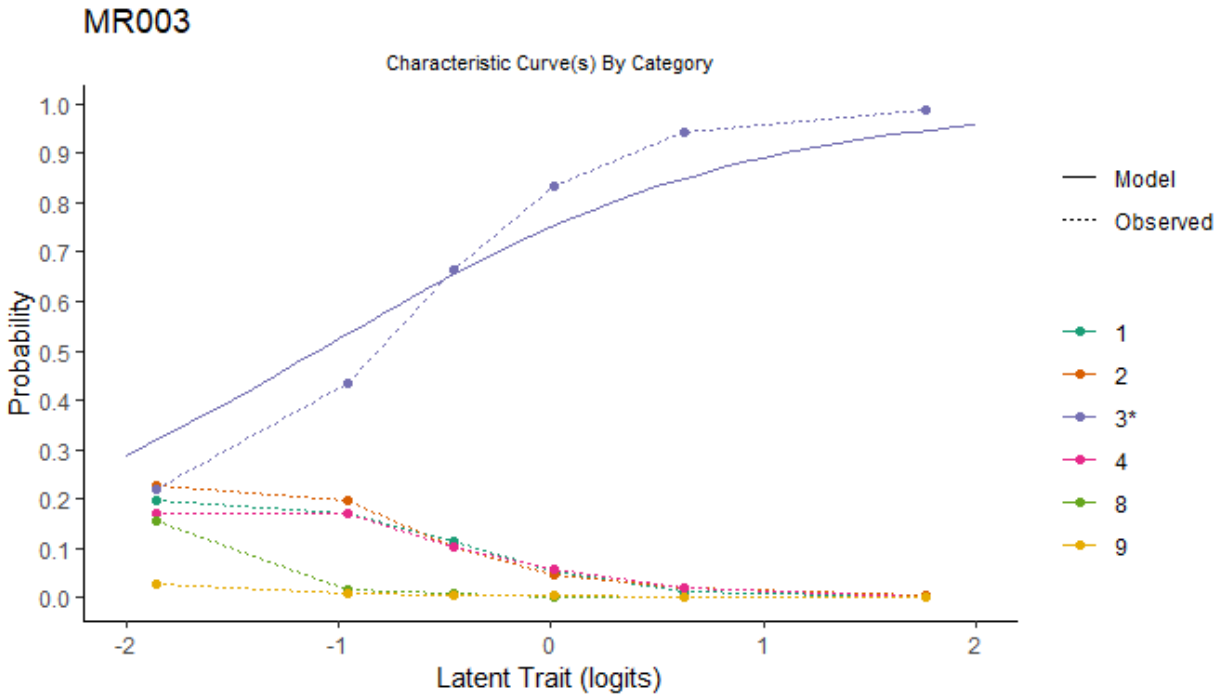
MR001



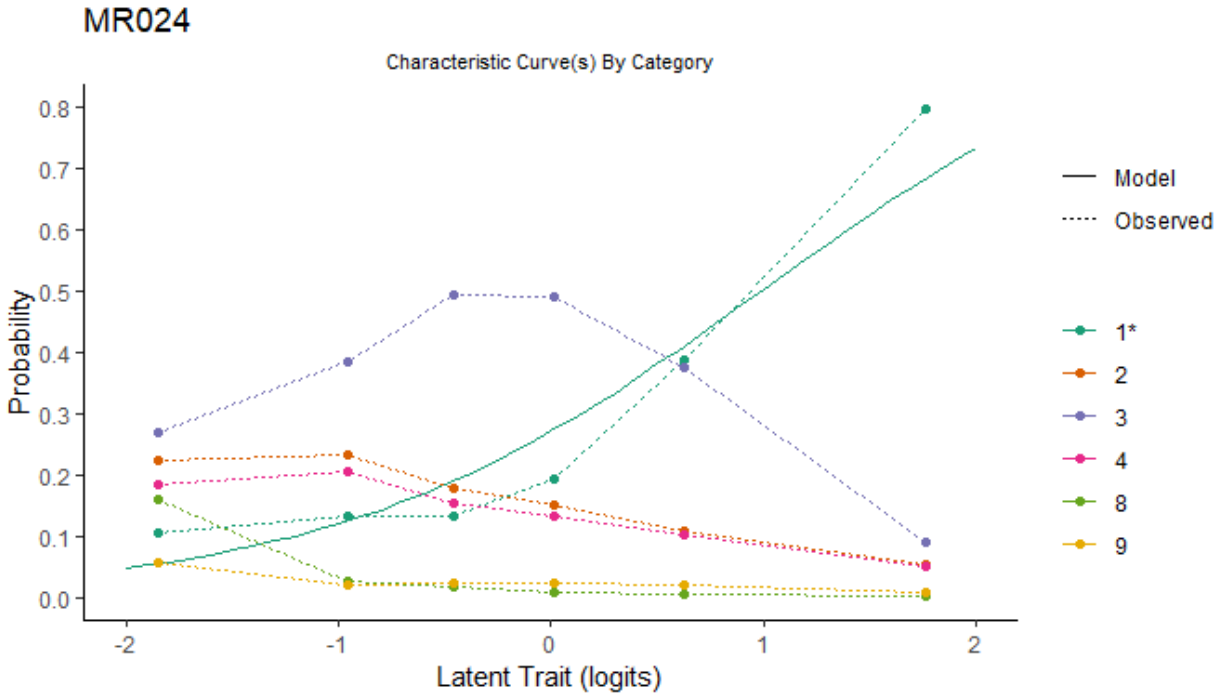
MR002



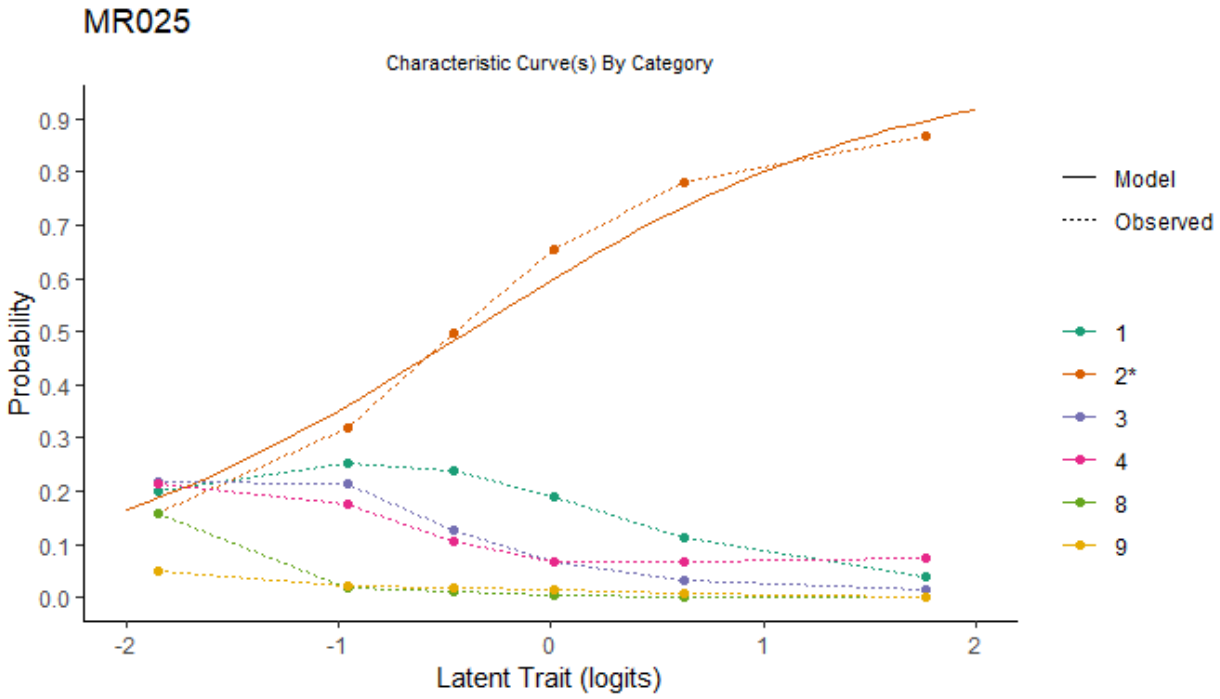
MR003


MR003
Characteristic Curve(s) By Category
Probability
Latent Trait (logits)
Model
Observed
1
2
3*
4
8
9

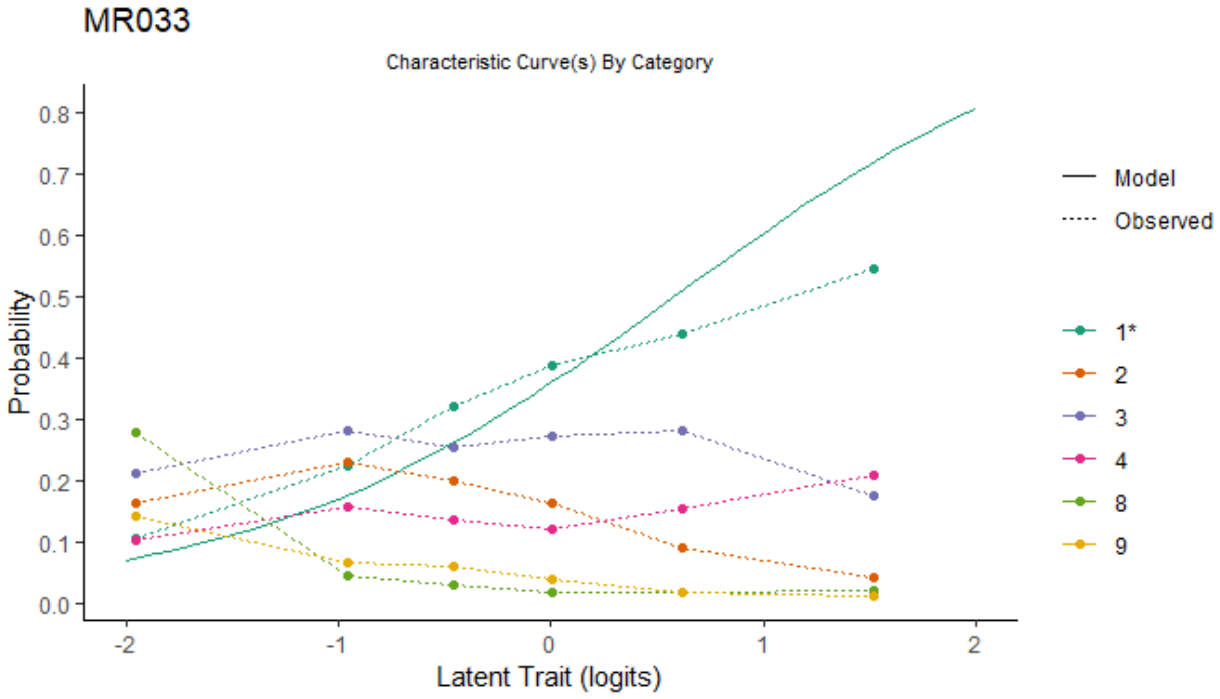
MR024


MR024
Characteristic Curve(s) By Category
Probability
Latent Trait (logits)
Model
Observed
1*
2
3
4
8
9

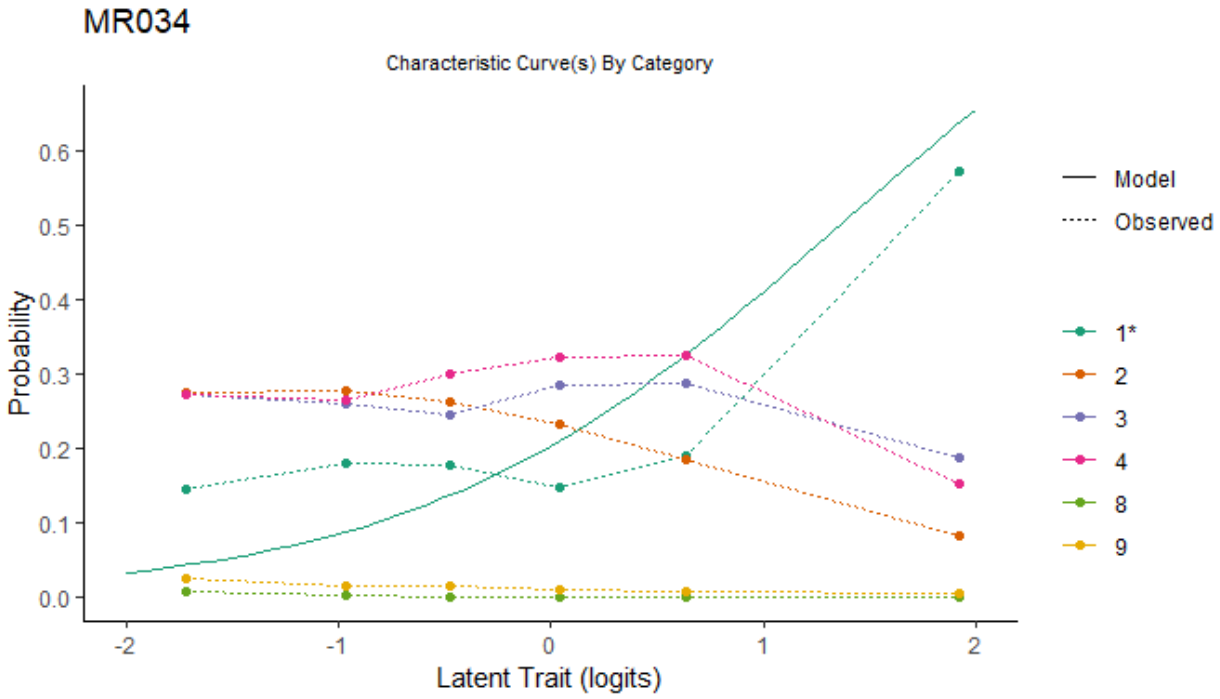
MR025


MR025
Characteristic Curve(s) By Category
Probability
Latent Trait (logits)
Model
Observed
1
2*
3
4
8
9

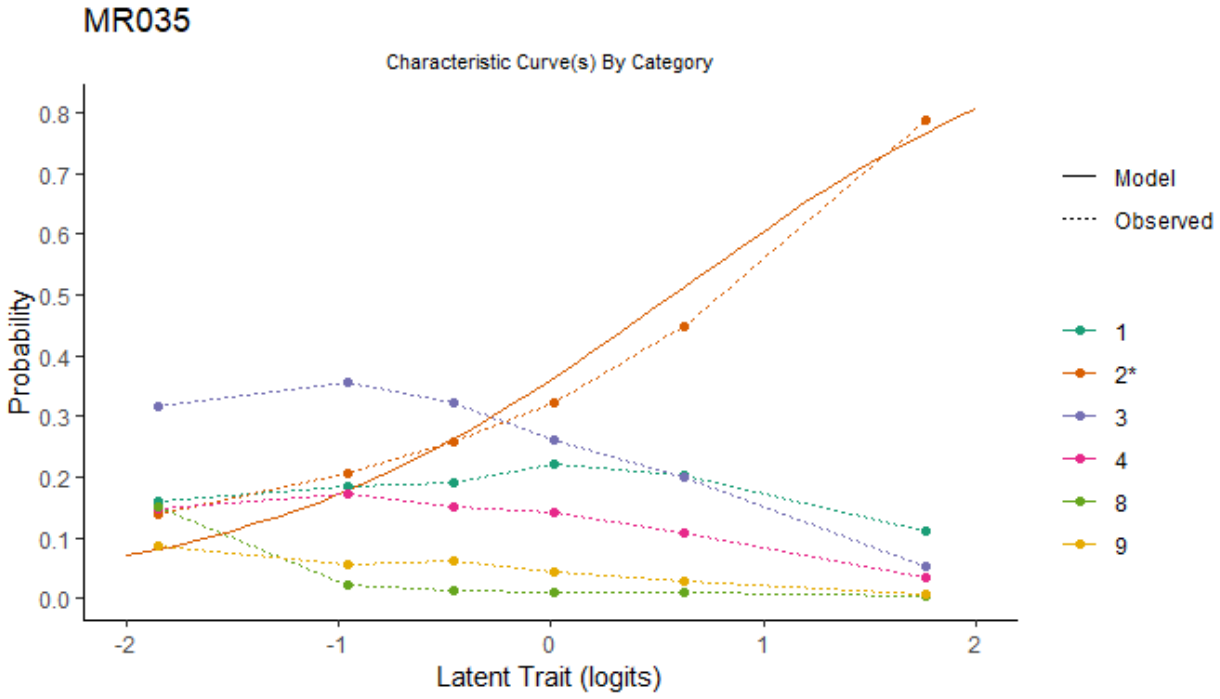
MR033


MR033
Characteristic Curve(s) By Category
Probability
Latent Trait (logits)
Model
Observed
1*
2
3
4
8
9

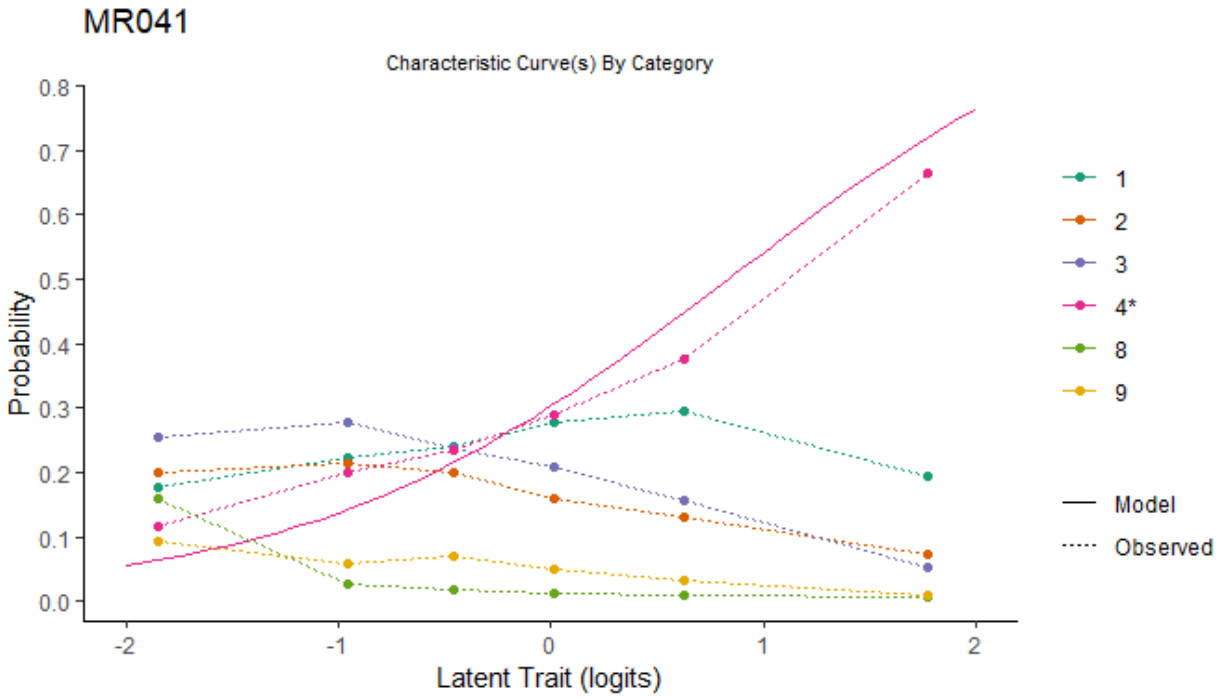
MR034


MR034
Characteristic Curve(s) By Category
Probability
Latent Trait (logits)
Model
Observed
1*
2
3
4
8
9

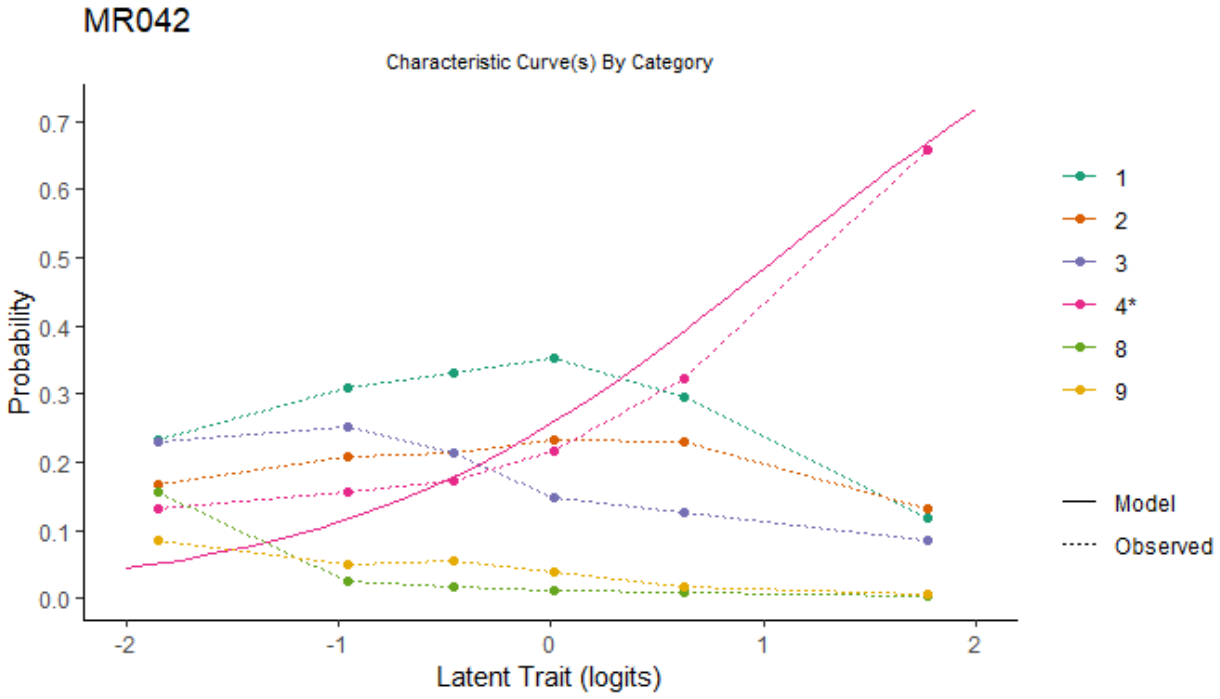
MR035


MR035
Characteristic Curve(s) By Category
Probability
Latent Trait (logits)
Model
Observed
1
2*
3
4
8
9

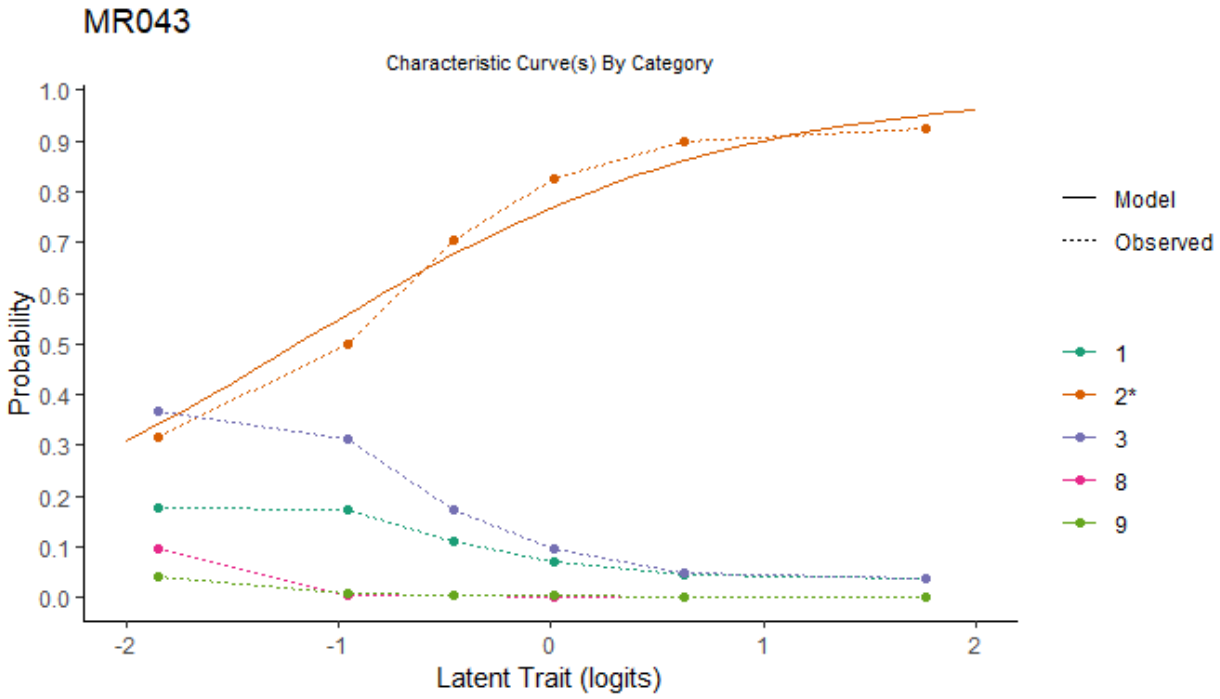
MR041


MR041
Characteristic Curve(s) By Category
Probability
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
-2
-1
0
1
2
Latent Trait (logits)
1
2
3
4*
8
9
Model
Observed

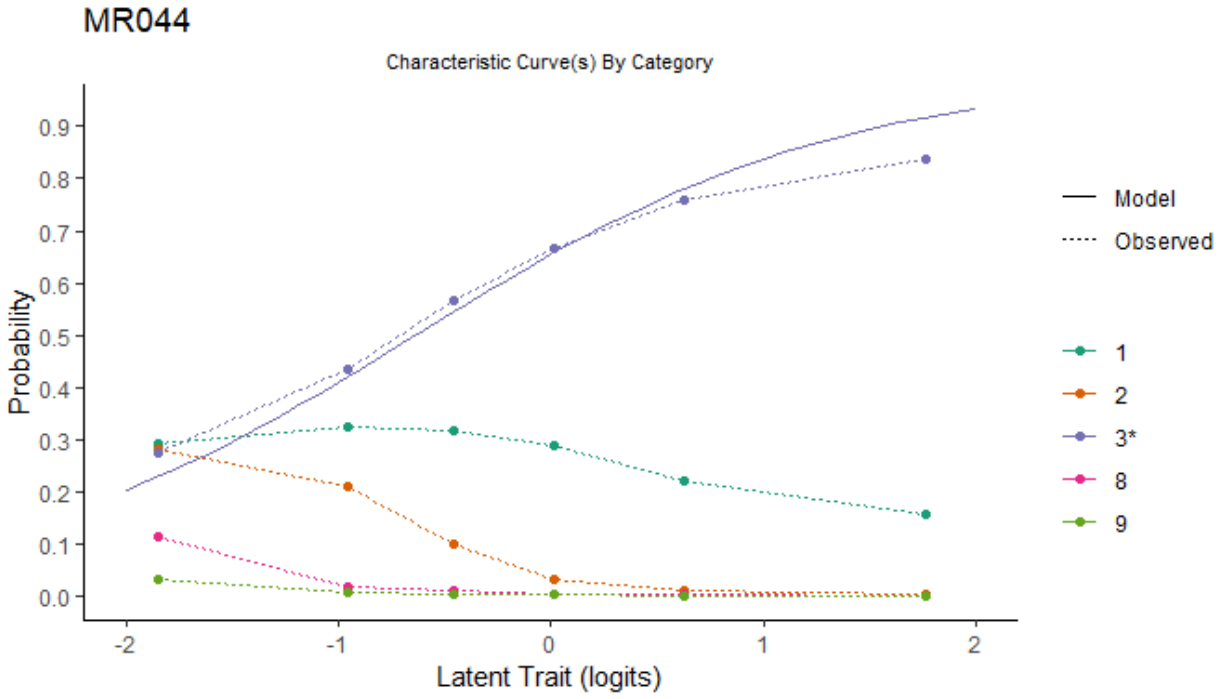
MR042


MR042
Characteristic Curve(s) By Category
Probability
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
-2
-1
0
1
2
Latent Trait (logits)
1
2
3
4*
8
9
Model
Observed

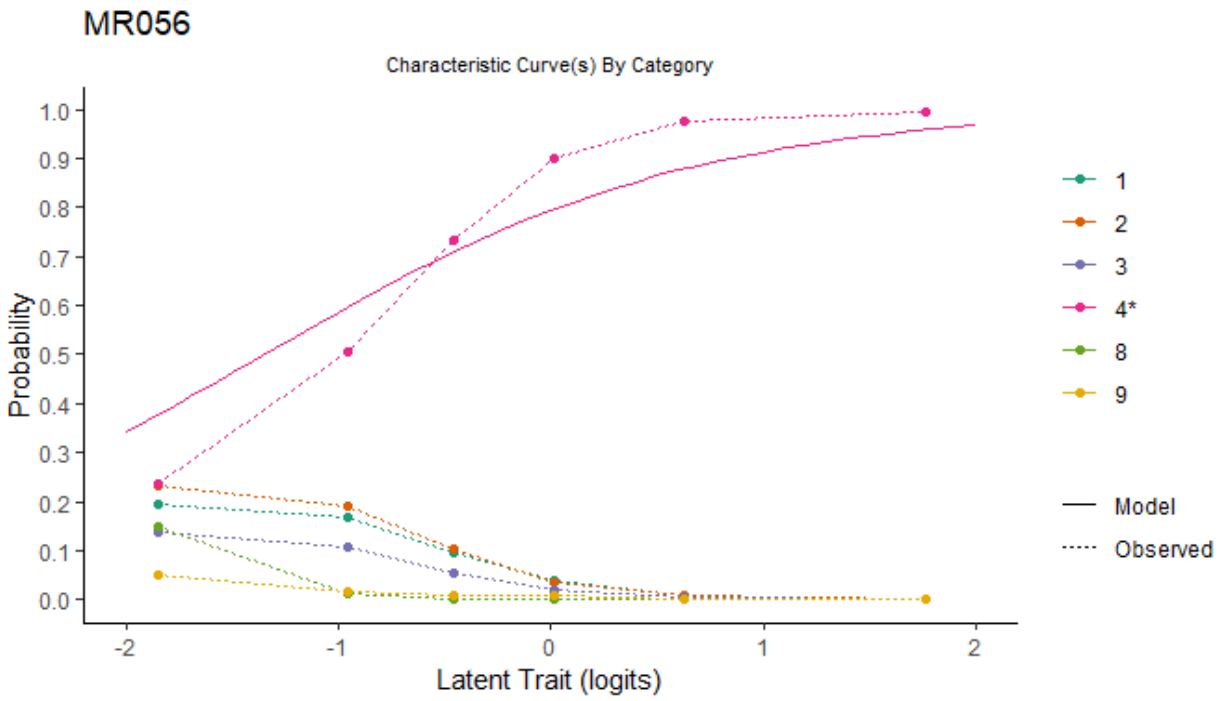
MR043


MR043
Characteristic Curve(s) By Category
Probability
Latent Trait (logits)
Model
Observed
1
2*
3
8
9

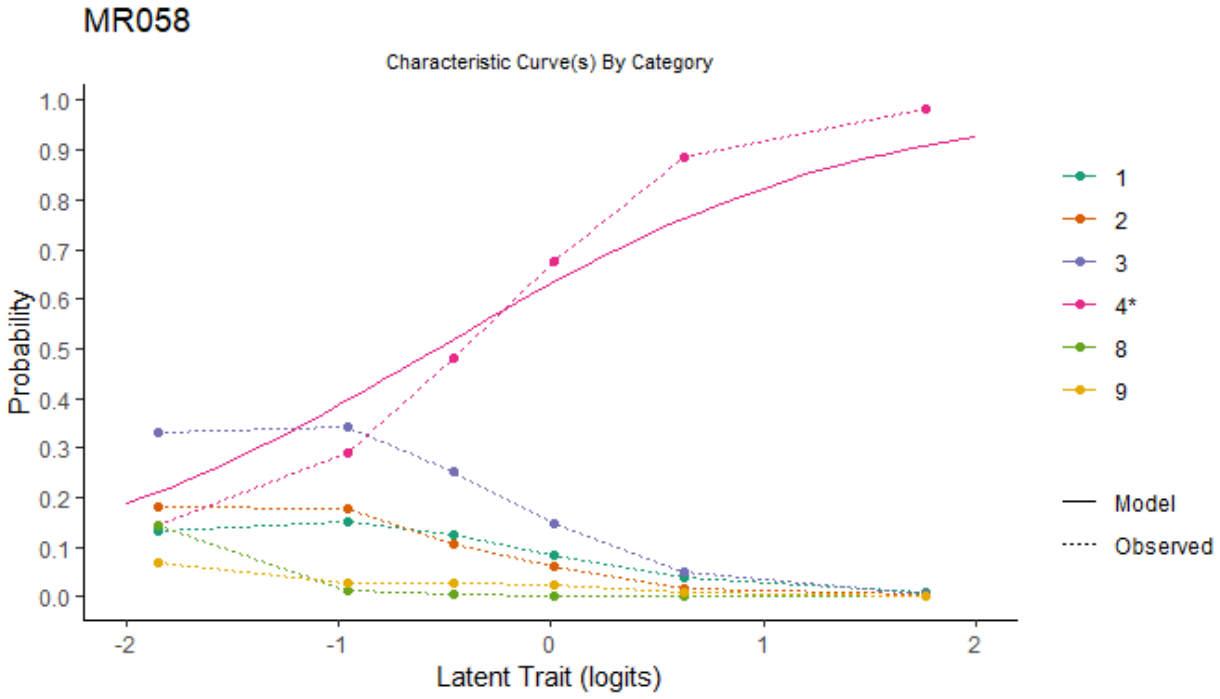
MR044


MR044
Characteristic Curve(s) By Category
Probability
Latent Trait (logits)
Model
Observed
1
2
3*
8
9

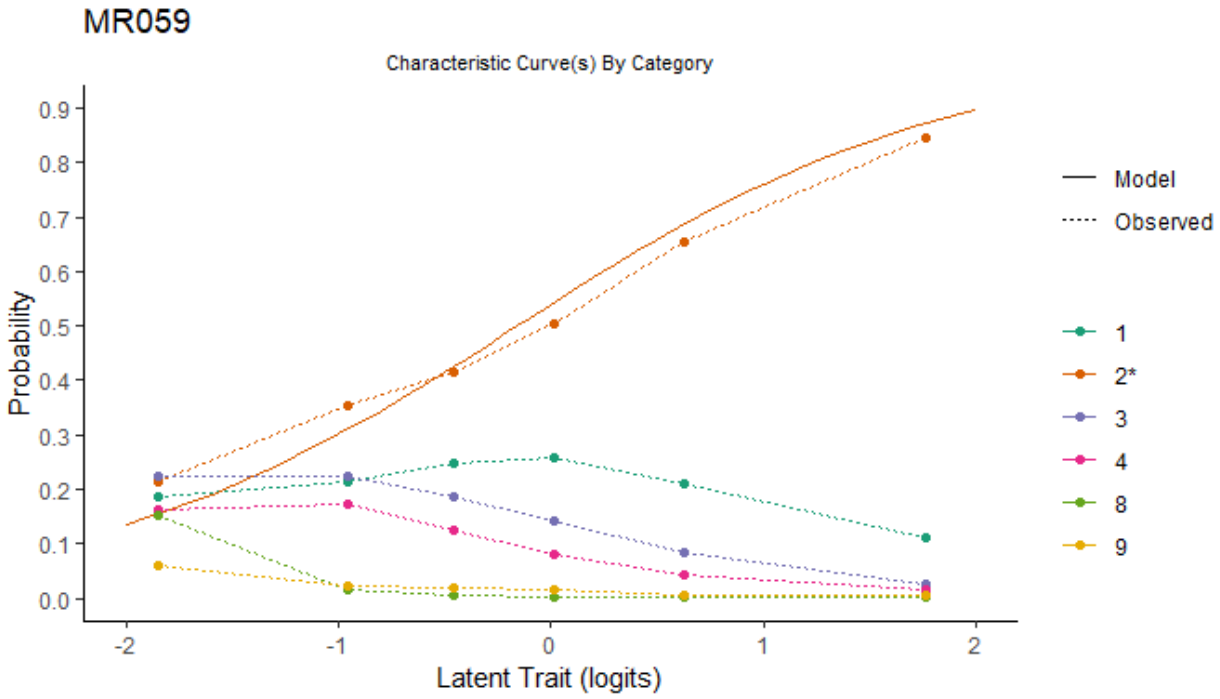
MR056


MR056
Characteristic Curve(s) By Category
Probability
Latent Trait (logits)
1
2
3
4*
8
9
Model
Observed

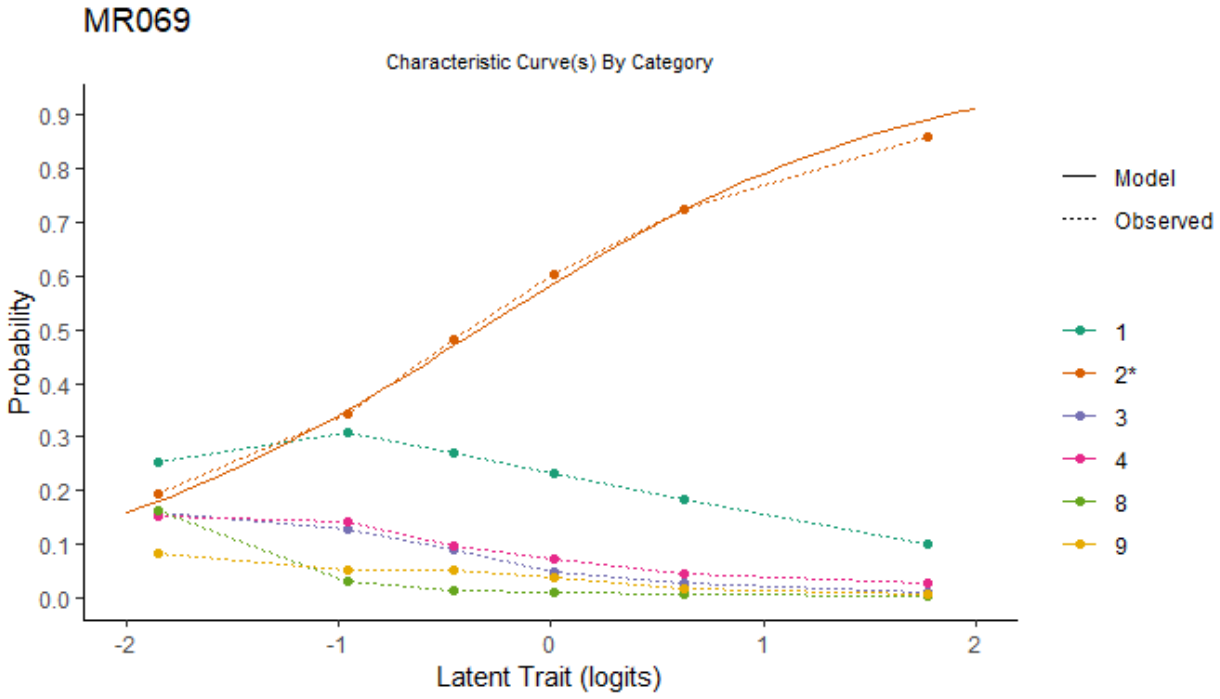
MR058


MR058
Characteristic Curve(s) By Category
Probability
Latent Trait (logits)
1
2
3
4*
8
9
Model
Observed

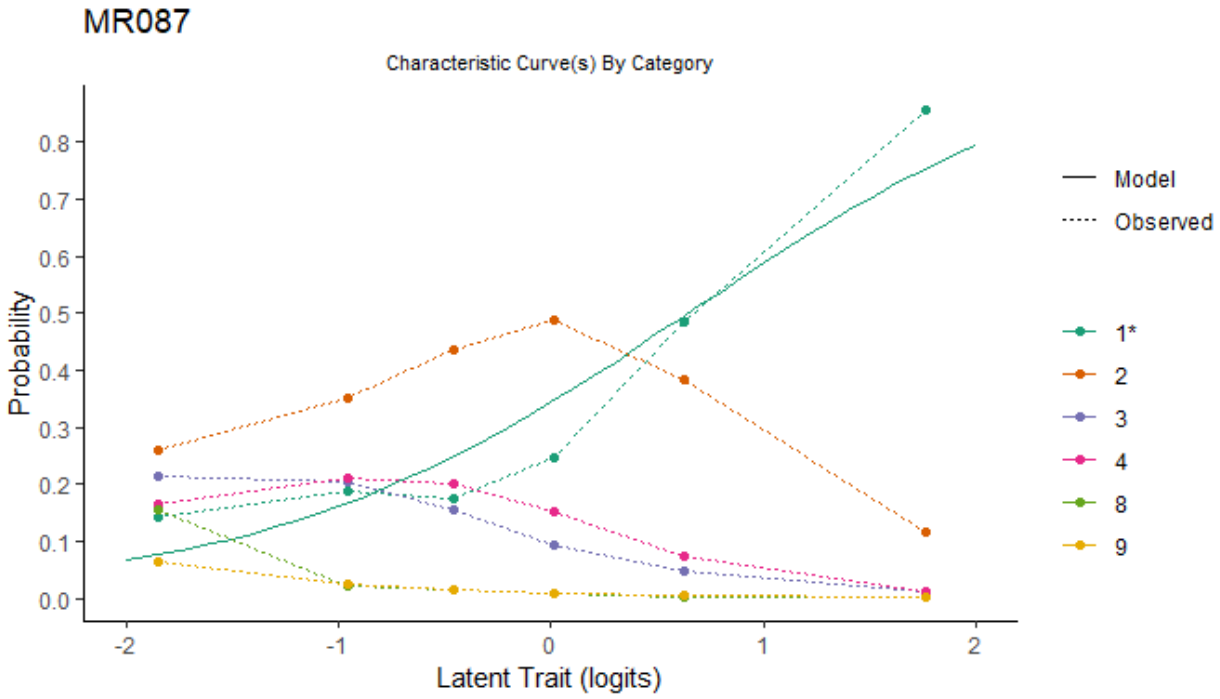
MR059


MR059
Characteristic Curve(s) By Category
Probability
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
-2
-1
0
1
2
Latent Trait (logits)
Model
Observed
1
2*
3
4
8
9

MR069


MR069
Characteristic Curve(s) By Category
Probability
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
-2
-1
0
1
2
Latent Trait (logits)
Model
Observed
1
2*
3
4
8
9

MR087


MR087
Characteristic Curve(s) By Category
Probability
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
-2
-1
0
1
2
Latent Trait (logits)
Model
Observed
1*
2
3
4
8
9