Renaissance

Applying Multidimensional Item Response Theory to Renaissance Assessments to Enhance Diagnostic Reporting Capacity

By William P. Skorupski, Associate Professor of Research, Evaluation, Measurement, and Statistics, School of Education, University of Kansas

Renaissance
PO Box 8036
Wisconsin Rapids, WI 54495-8036
(800) 338-4204
www.renaissance.com

10/17

Contents

4 Executive Summary
5 Overview of Multidimensional Item Response Theory
9 Analyses for Renaissance Star Reading® and Renaissance Star Math®
9 Results for Math and Reading
12 Linking and Equating
13 Final Steps
16 References

Figures

6 Figure 1. Graphic representation of the MIRT bi-factor model for a hypothetical 18-item test measuring three Skill Areas
7 Figure 2. Example MIRT item
8 Figure 3. Marginal item performance by Skill Area

Table

10 Table 1R. Number of retained items, Skill Areas, and examinees analyzed at each grade level (includes multiple duplicate items across grade levels) for Star Reading
10 Table 1M. Number of retained items, Skill Areas, and examinees analyzed at each grade level (includes multiple duplicate items across grade levels) for Star Math
11 Table 2R. Deviance Information Criterion (DIC) statistics by grade level and type of calibration for Star Reading analyses
11 Table 2M. Deviance Information Criterion (DIC) statistics by grade level and type of calibration for Star Math analyses
14 Table 3R. List of all Skill Areas and number of items per Skill Area for Star Reading MIRT analyses
15 Table 3M. List of all Skill Areas and number of items per Skill Area for Star Math MIRT analyses

Executive Summary

The goal of the Multidimensional Item Response Theory (MIRT) project for Renaissance is to increase the diagnostic reporting capacity for Renaissance assessments. The proposed methodology for meeting this goal was to evaluate the dimensionality of Renaissance Star items with respect to the learning progressions in math and reading. The learning progressions create testable hypotheses about the dimensional structure of the Star assessment items. This dimensionality was evaluated via MIRT models implied by the interrelated Skill Areas suggested by the learning progressions. Multidimensionality in the data was previously established by examining residual correlations from unidimensional Item Response Theory (IRT) calibrations of data. These residual correlations suggested that significant dimensionality was present in the data. After multidimensionality had been established, bi-factor MIRT models suggested by the Skill Area coding of items were fit to the data. These MIRT models showed significant improvement in model-data fit, thus validating the dimensionality hypotheses. Lastly, separate calibrations of items at different grade levels were linked to a common metric. This allows administration of items from across grade levels while still being able to produce scores on a constant scale. This report summarizes analyses for math and reading, gives insight into the meaning of Skill Areas, and demonstrates the added value of using MIRT for scaling.

Dr. William P. Skorupski is an Associate Professor of Research, Evaluation, Measurement, and Statistics in the School of Education at the University of Kansas. He teaches courses in item response theory (IRT), classical test theory, computer programming, ANOVA, regression, and multivariate statistics. His research focuses on applications of IRT, the evaluation of standard setting processes, the use of Bayesian statistics for solving practical measurement problems, and the implementation and estimation of innovative measurement models. Dr. Skorupski received his Ed.D. in Psychometric Methods from the University of Massachusetts Amherst in 2004. Dr. Skorupski specializes in item response theory and applications, psychometric methods, scaling and test score equating.

Overview of Multidimensional Item Response Theory

Multidimensional Item Response Theory (MIRT; Reckase, 2009) is a family of statistical models used for scaling assessments that measure multiple traits at a time. As the name implies, it is the multidimensional extension of unidimensional Item Response Theory (IRT). The MIRT approach to scaling differs from traditional IRT in that it allows one to model multiple skills simultaneously. Thus, data may be considered in much more complex ways. The two-parameter logistic (2-PL) MIRT model relates the probability of success on a dichotomously scored item as a function of a vector of latent traits:

$$P(u_{ij} = 1 \mid \underline{\theta}_i) = \frac{e^{\underline{a}_j \underline{\theta}_i' + d_j}}{1 + e^{\underline{a}_j \underline{\theta}_i' + d_j}}$$

where u_{ij} is the scored response to item j by person i, $\underline{\boldsymbol{a}}_j$ is a vector of discrimination or slope parameters for item j, d_j is the intercept or 'easiness' parameter for item j, and $\underline{\theta}_i'$ is a transposed vector of latent trait parameters for person i. This model is the multidimensional extension of the unidimensional 2-PL IRT model for dichotomously scored items (Birnbaum, 1968):

$$P(u_{ij} = 1 \mid \underline{\theta}_i) = \frac{e^{a_j \theta_i + d_j}}{1 + e^{a_j \theta_i + d_j}}.$$

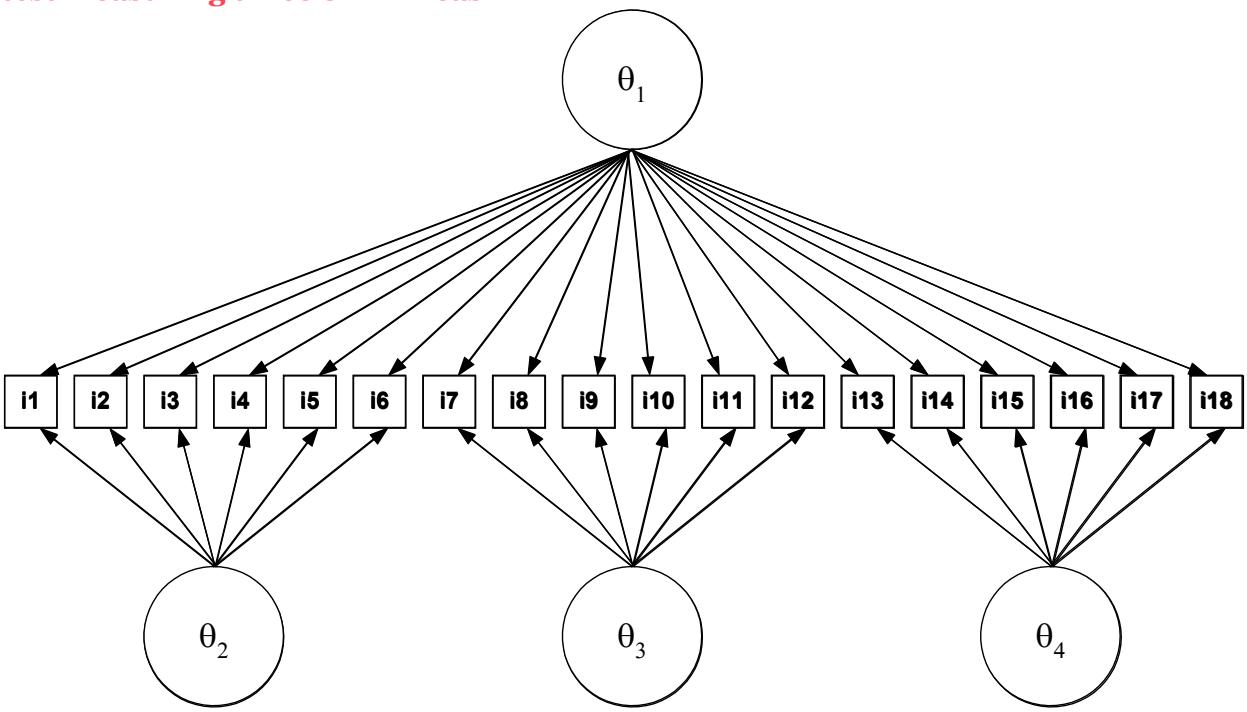
The unidimensional 2-PL IRT model only has one *a*-parameter per item (i.e., there is only one trait being measured) and thus there is only one person effect (θ) in the model as well.

Renaissance assessment data were analyzed by a particular kind of MIRT model referred to as a "bi-factor model." Bi-factor models were first introduced by Holzinger and Swineford (1937) with linear confirmatory factor analysis (CFA), but the same concepts may be applied to MIRT (which is essentially a nonlinear CFA). In a bi-factor model, many dimensions may be represented, but a given item measures two and only two dimensions; the overall dimension and one Skill Area dimension. If, for example, 30 Skill Areas were represented in the pool of Star items, the entire MIRT model would represent 31 dimensions (an "overall" dimension plus one additional dimension per Skill Area). However, any individual item only measures two dimensions, overall and Skill Area specific. All other *a*-parameters (discrimination) from the other dimensions are constrained to be equal to zero. In this way, a bi-factor approach to estimation ensures that the overall dimension and all Skill Area dimensions are orthogonal (statistically independent and therefore uncorrelated with each other). For reading or math, the "overall" dimension represents all of the interconnectedness that exists across these constructs. The Skill Area dimensions therefore represent only the uniqueness of those Skill Areas within the construct. Thus, the bi-factor model is perfectly suited for diagnostic purposes.

Figure 1 contains a graphic representation of this bi-factor dimensional structure. For this hypothetical 18-item test, all 18 items are influenced by the θ_1 dimension, indicating the overall/global trait. Items 1–6 are influenced by the θ_2 dimension, indicating the first Skill Area dimension, items 7–12 are influenced

by the θ_3 dimension, indicating the second Skill Area dimension, and items 13–18 are influenced by the θ_4 dimension, indicating the third Skill Area dimension.

Figure 1. Graphic representation of the MIRT bi-factor model for a hypothetical 18-item test measuring three Skill Areas



This MIRT model is referred to as "compensatory" because trait values (θ) across dimensions are weighted (by their respective *a*-parameters) and then added together when evaluating the probability of a category response. Thus, a relatively low value for one trait may be compensated by a relatively high value on another trait, provided the *a*-parameter for that trait is relatively large. Because the bi-factor structure ensures that every item only measures two traits, we can re-express the model this way:

$$P(u_{ij}=1\,|\,\underline{\theta}_i)=\frac{e^{\underline{a}_j\underline{\theta}_i'+d_j}}{1+e^{\underline{a}_j\underline{\theta}_i'+d_j}}=\frac{e^{a_{jG}\theta_{iG}+a_{jS}\theta_{iS}+d_j}}{1+e^{a_{jG}\theta_{iG}+a_{jS}\theta_{iS}+d_j}}$$

where $\underline{a}_{jG}$ is the *a*-parameter for the "overall" dimension, $\underline{a}_{jS}$ is the *a*-parameter for the Skill Area dimension, and all other values are defined as per the previous equation. Figure 2 contains a graphic representation of a MIRT item response function for an item measuring two latent traits.

Figure 2. Example MIRT item

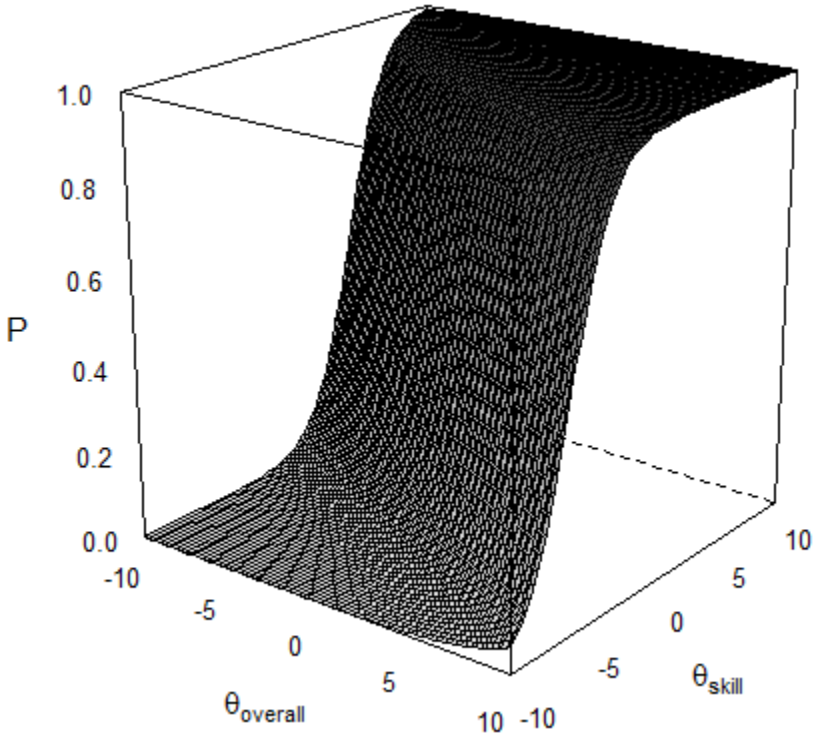
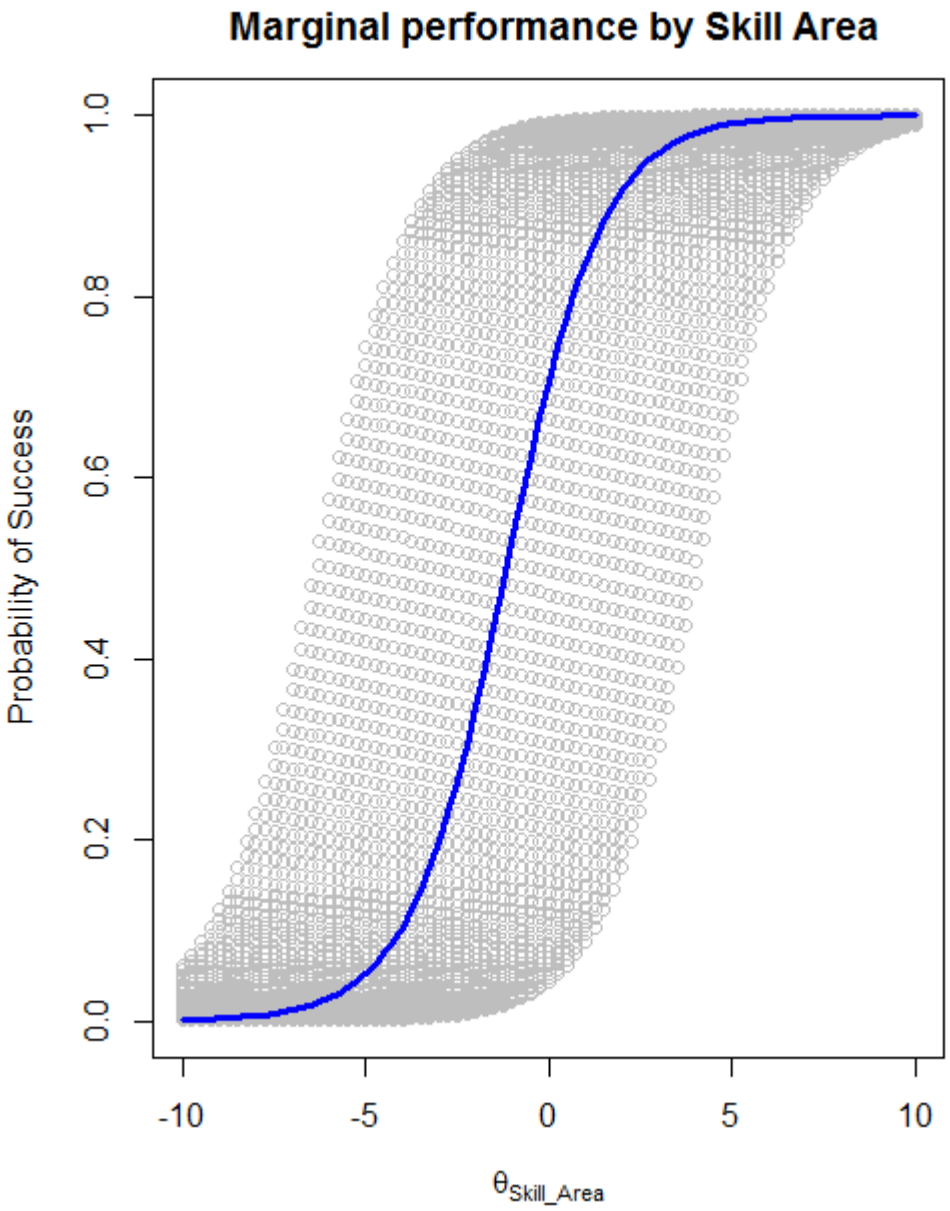


Figure 3 contains a graphic depiction of the marginal performance by Skill Area on the same item as shown in figure 2. Marginal performance here refers to performance based on Skill Area but averaged over the overall dimension. Thus, the calculations in Figure 3 represent what the item can tell us about an examinee's Skill Area score, without respect to his/her overall reading performance.

Figure 3. Marginal item performance by Skill Area



The graphic in figure 3 shows how well the example item discriminates along its Skill Area dimension. The vertical range of dots at each θ_{Skill_Area} value around the dark line shows the change in the probability of success for a range of overall ability levels.

Analyses for Renaissance Star Reading and Renaissance Star Math

For the Star MIRT analyses, the Renaissance learning progressions were used to create testable hypotheses about how skills relate to each and how distinct they are from one another. Essentially, dimensionality in a MIRT model is inferred from the presence of highly intercorrelated pathways as evidence of similar traits and relatively uncorrelated pathways as evidence of distinct traits. It is expected that all items measure something common (either reading or math). But the purpose of the MIRT analyses was to evaluate any extra-dimensionality associated with skill-specific sources of variability. If skill-specific dimensionality is present in the data, then information from students' item responses could be used to infer not just overall ability in reading or math, but also skill-specific ability, which would be very beneficial for diagnostic purposes.

Dimensionality was established by evaluating the relative fit of various confirmatory MIRT models to the same data. Initial unidimensional IRT calibrations showed that the items all measured a common trait, but also that items grouped within Skill Areas shared nontrivial residual correlations. Such residual correlations suggest that significant dimensionality was present in the data. Further analyses of the Star data in reading and math first considered dimensionality suggested by the learning progressions. In the learning progressions, every item measures one fairly specific Skill. Skills are further clustered within Skill Areas, organizational units of understanding which have courser grain sizes than skills. Analyses showed that skills were too fine a grain size to be used profitably for MIRT calibration, but Skill Areas worked perfectly. Bi-factor MIRT calibrations that coded items by skills produced too many dimensions, with too few items per dimension; many calibrations failed to converge. However, coding items by Skill Areas produced a manageable number of dimensions, with several items per dimension being measured. All these calibrations converged and showed significant improvement in model-data fit compared to unidimensional calibrations.

For the final set of calibrations, each unique Skill Area represented at a grade level was mapped onto one and only one MIRT dimension. The dimensionality of each assessment is evaluated as N(Skill Areas) + 1, an "overall" dimension to represent the intercorrelated aspects of the construct (e.g., "Reading" or "Math"), plus one dimension per Skill Area, which represents the unique, independent dimensionality of that Skill Area within the construct.

Results for Math and Reading

Data from Renaissance Star Assessments were analyzed using the *mirt* package in R (Chalmers, 2012). There may be hundreds or even over a thousand items administered at a grade level, but most examinees only take a small handful. For every grade level and subject, the person-by-item data matrix was always analyzed twice: (1) a unidimensional 2-PL IRT model (basically, the MIRT model, but with only an overall trait per item, no Skill Area dimensions) was fit to the data, and (2) the bi-factor MIRT model with an overall dimension plus one additional dimension per Skill Area was fit to the same data.

Global model-fit comparisons tests known as “Deviance Information Criterion” statistics (DIC; Spiegelhalter, Best, & Carlin, 1998) were then conducted to demonstrate that the multidimensional solution explained the data significantly better. The DIC statistic represents how well the model explains the data, plus a penalty factor for model complexity; thus a reduction in DIC shows improved model-data fit. For all calibrations, the MIRT model was a significant improvement in model-data fit. The data analyzed for the initial calibrations are summarized in Table 1R (Star Reading) and Table 1M (Star Math). Summaries of DIC statistics are contained in Table 2R (Star Reading) and Table 2M (Star Math).

Table 1R. Number of retained items, Skill Areas, and examinees analyzed at each grade level (Includes multiple duplicate items across grade levels) for Star Reading

Grade level	N Items	N Skill Areas	N Examinees
1	253	9	23235
2	646	17	31980
3	1191	19	36959
4	786	20	35783
5	769	24	34269
6	343	22	27883
7	848	24	40715
8	719	24	39071
9	118	23	18284
10	151	23	21288
11	83	21	14334
12	46	18	8921

Table 1M. Number of retained items, Skill Areas, and examinees analyzed at each grade level (Includes multiple duplicate items across grade levels) for Star Math

Grade level	N Items	N Skill Areas	N Examinees
1	877	15	24145
2	914	16	35931
3	1128	19	39638
4	445	22	40472
5	894	25	38532
6	641	26	34162
7	337	25	30902
8	256	25	31172
9	439	28	24538
10	343	26	22571
11	118	23	15852
12	40	16	9636

Table 2R. Deviance Information Criterion (DIC) statistics by grade level and type of calibration for Star Reading analyses

Grade level	DIC from unidimensional calibration (IRT)	DIC from multidimensional calibration (MIRT)	Evidence for multidimensionality DIC[MIRT] < DIC[IRT]
1	1473894	1470367	✓
2	2079129	2073996	✓
3	3036820	3025883	✓
4	2122232	2114046	✓
5	1981830	1974203	✓
6	930907	928974	✓
7	2456295	2448736	✓
8	2301826	2295688	✓
9	371103	370740	✓
10	544069	543564	✓
11	301036	300786	✓
12	143042	142916	✓

Table 2M. Deviance Information Criterion (DIC) statistics by grade level and type of calibration for Star Math analyses

Grade level	DIC from unidimensional calibration (IRT)	DIC from multidimensional calibration (MIRT)	Evidence for multidimensionality DIC[MIRT] < DIC[IRT]
1	2441600	2417509	✓
2	3221319	3188589	✓
3	3573697	3544116	✓
4	1798496	1785646	✓
5	2363343	2343066	✓
6	1587175	1575782	✓
7	807566	803829	✓
8	594472	592219	✓
9	1387108	1373313	✓
10	1004780	997404	✓
11	304331	302918	✓
12	100395	100067	✓

Linking and Equating

For each subject, after successfully calibrating all items with MIRT, and demonstrating that these solutions were superior to unidimensional calibrations, one additional task remained: creating a common scale across the 12 sets of parameter estimates. Thus, item parameter estimates from separate calibrations need to be linked to one equated common metric. The technical details for the linking procedures are described in other Renaissance internal documents, but this report will provide a broad overview.

In MIRT calibrations, the latent traits (θ) have no natural scale, because they are not directly observed. Consequently, any scale can be used to interpret item parameters, but one *must* be chosen in order to identify the model (without choosing a scale, a literally infinite number of possible solutions would fit the data equally well by simply changing the scale). When items have been calibrated with MIRT across separate datasets, their scales must be linked to a common, equated metric before comparisons of item parameters can be made. Likewise, if constructing test forms using items from across grade levels, it is essential that this linking and equating have been conducted.

The scale of item parameters is determined by the properties of the items themselves as well as the distribution of ability in the examinee sample used for calibration. MIRT item parameters are said to be *scale invariant*, however, because they are assumed to be independent of the particular ability distribution of respondents. That is, parameter estimates from calibrations of the same items taken by different examinees will be equivalent within a linear transformation. This linear transformation accounts for the differences in the ability distributions between the groups of examinees, and thereby places their respective values on a common scale. If item parameters are invariant, a scatterplot of common-item intercept (d) values between calibrations will show these parameters falling on or near a straight line. This line represents the slope and intercept for the equating transformation. It was observed that strong item parameter invariance was present, because *d*-values (intercepts from the MIRT model) from items appearing at adjacent grade levels were always very strongly linearly related.

The adaptive administration algorithm used in all Star Assessments means that items are frequently administered across grade levels. This works well for our needs, because it means we have some common items being administered to different groups (these items are often referred to as “anchor items” in this context, because they are used to hold the scale in place). If their parameters were estimated similarly across calibrations, then they would be already on a common metric. However, if the parameters of anchor items systematically vary over calibrations, it tells us something about differences in the examinees used to calibrate the data. If those groups of examinees differ in ability (which one would expect, since the groups represent grade levels), then the only way to infer those differences is to see how the item parameter estimates for common items change.

To link all items for a subject onto a common metric, MIRT item parameter estimates were collected from across all 12 calibrations (one for each grade level within the set of reading or math calibrations). This includes duplicate items which have been administered at more than one grade level. Since

hundreds of items are calibrated at a grade level, it is reasonable to find that not all items have good statistical properties at every (or perhaps at any) grade level. Thus, the next analytic step was to remove any item at a grade level with *a*-parameters (on either dimension) less than 0.1. This step ensured that every remaining item was positively discriminating along *both* of its dimensions (because dimension effects are independent and additive, sometimes an item could positively discriminate globally, but have a close-to-zero or even negative *a*-parameter on one of its dimensions). The adoption of the 0.1 criterion was agreed upon by Renaissance staff to maximize inclusion of items in the item pool, while ensuring that every retained item demonstrated high psychometric quality for diagnostic purposes.

For each subject, 12 separate calibrations were conducted across grade levels. One calibration is chosen as the baseline metric, and then all other calibrations are linked to that scale. A middle grade level (6th) was chosen as the baseline so that the maximum distance between any two equating steps would be minimized. Equating analyses were conducted in a series of 11 steps, with each adjacent grade level being linked to the baseline metric before the next grade level results were linked to the next one.

Final Steps

The last step in the MIRT calibration process was to select the final set of items for inclusion in an item bank. Items were often administered across multiple grade levels, but the item bank should only contain one entry per item. Item parameters were selected for inclusion by identifying the largest Skill Area *a*-parameter per item after equating was complete. For any item administered at multiple grade levels, the combination of parameters chosen to represent that item would be the one for which its Skill Area *a*-parameter was largest. Additionally, the item was "tagged" at that grade level, to indicate that the item was maximally discriminating at that grade. If an item only appeared at one grade level, then its equated item parameters were simply retained and it was tagged at that grade level.

For reading, 2,278 unique items, representing 26 Skill Areas and 12 grade levels were retained from the MIRT analyses. For math, 2,106 unique items, representing 33 Skill Areas and 12 grade levels were retained. Note that more Skill Areas were identified and numbered by Renaissance content staff, but not all these were represented in the data, or retained during/after calibration. A summary of items by Skill Areas for reading and math are contained in Table 3R and Table 3M, respectively.

Table 3R. List of all Skill Areas and number of items per Skill Area for Star Reading MIRT analyses

#	Skill Area	N Items
1	Analysis and Comparison	1
2	Argumentation	187
3	Author's Purpose and Perspective	14
4	Author's Word Choice and Figurative Language	141
5	Cause and Effect	11
6	Character and Plot	83
7	Compare and Contrast	24
8	Connotation	2
9	Context Clues	68
10	Conventions and Range of Reading	77
11	Figures of Speech	15
12	Inference and Evidence	14
13	Main Idea and Details	68
14	Multiple-Meaning Words	62
15	Point of View	40
16	Prediction	22
17	Sequence	19
18	Setting	37
19	Structural Analysis	77
20	Structure and Organization	28
21	Summary	17
22	Synonyms and Antonyms	89
23	Text Features	3
24	Theme	44
25	Vocabulary in Context	1113
26	Word Relationships	22

Table 3M. List of all Skill Areas and number of items per Skill Area for Star Math MIRT analyses

#	Skill Area	N Items
1	Algebraic Thinking	42
2	Angles, Segments, and Lines	58
3	Combinatorics and Probability	8
4	Congruence and Similarity	21
5	Coordinate Geometry	9
6	Data Representation and Analysis	165
7	Decimal Concepts and Operations	145
8	Fraction Concepts and Operations	104
9	Geometry: Three-Dimensional Shapes and Attributes	13
10	Geometry: Two-Dimensional Shapes and Attributes	38
11	Linear Expressions, Equations, and Inequalities	94
12	Matrices, Vectors, and Complex Numbers	1
13	Measurement	73
14	Money and Time	129
15	Nonlinear Expressions, Equations, and Inequalities	2
16	Numerical and Variable Expressions	27
17	Patterns, Sequences, and Series	77
18	Percents, Ratios, and Proportions	34
19	Perimeter, Circumference, and Area	50
20	Polygons	14
21	Polynomial Expressions and Functions	9
22	Positive and Negative Rational Numbers	11
23	Powers, Roots, and Radicals	22
24	Quadratic Expressions, Equations, and Inequalities	6
25	Relations and Functions	3
26	Right Triangles and Trigonometry	4
27	Surface Area and Volume	12
28	Systems of Equations and Inequalities	6
29	Transformations	12
30	Whole Numbers: Addition and Subtraction	300
31	Whole Numbers: Counting, Comparing, and Ordering	150
32	Whole Numbers: Multiplication and Division	210
33	Whole Numbers: Place Value	257

References

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick (Eds.), *Statistical theories of mental test scores* (pp. 397–479). Addison-Wesley.

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

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

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

Spiegelhalter, D., Best, N. G., & Carlin, B. P. (1998). *Bayesian deviance, the effective number of parameters, and the comparison of arbitrarily complex models*. Research Report 98-009*,* Division of Biostatistics, University of Minnesota*.*