A Comparison of Item Parameter Estimates in Pychometrik and the Existing Item Calibration Tool

January 2021

Wei He, Psychometric Solutions
Emily Bo, Psychometric Solutions
Patrick Meyer, Psychometric Solutions
Russ Grandgeorge, Data Operations

© 2021 NWEA. NWEA and MAP Growth are registered trademarks of NWEA in the U.S. and in other countries. All rights reserved. No part of this document may be modified or further distributed without written permission from NWEA.

This report benefited from the expert copy editing from Kelly Rivard.

Suggested citation: He, W., Bo, E., Meyer, P., & Grandgeorge, R. (2020). *A comparison of item parameter estimates in Pychometrik and the existing item calibration tool*. NWEA.

Table of Contents

List of Tables

List of Figures

Executive Summary

The design features of the existing MAP® Growth™ item calibration system have become too restrictive to support the expanding business and product development needs of NWEA®. As such, Pychometrik, a new item calibration tool built with Python, was recently developed to allow psychometricians autonomy in the item calibration process for all NWEA shelf products. To evaluate the performance of Pychometrik against the existing tool, two studies were conducted: (1) an item parameter estimate comparability study using both real data and simulations and (2) an item parameter recovery simulation study to evaluate the parameter recovery accuracy of the new tool. Results from both studies support the use of Pychometrik for future item calibrations.

1. Introduction

The existing MAP® Growth™ item calibration tool has been in place for many years. However, although the tool has served the needs of MAP Growth item calibration well, its design features are no longer able to support the expanding business and product development needs of NWEA®. As such, Pychometrik, a new calibration tool built with Python, was developed to handle item calibration for all NWEA shelf products. The name *Pychometrik* is a play on two words. It combines ***Py***thon and psy***chometrik*** because it is a software tool for psychometrics written in the Python programming language. Thus, there is no "s" is Pychometrik.

To investigate the degree to which item parameter estimates from both the existing and new calibration tools are comparable to each other, two studies were conducted: (1) an item parameter estimate comparability study using both real data and simulations to evaluate the comparability of the item parameter estimates between the two tools and (2) an item parameter recovery simulation study to evaluate the parameter recovery accuracy of Pychometrik. Challenges and issues with the existing item calibration tool include the following:

1. It is designed specifically for Rasch-based items only using a fixed-person estimation design. It cannot support the estimation of other standard item response theory (IRT) models such as two-parameter logistic (2PL), three-parameter logistic (3PL), or polytomous models. It also does not support free estimation of item and person parameters. Every use of the system must include fixed-person parameters.
2. It is managed outside of the NWEA Psychometric Solutions team, removing psychometricians from the process. This organization makes it challenging for psychometricians to control the item calibration process, a key psychometric practice, and stay informed of or catch problems that could occur in item calibration.
3. It is nearly impossible for differential item functioning (DIF) or item parameter drift studies to be conducted within the existing tool. Equally challenging is to conduct quality assurance (QA) checks and replication.
4. It can barely support the use of a targeted field test sample for item calibration. A frequent issue in item calibration is that items are calibrated with dominant responses from students who have never been exposed to the curriculum standards to which the items are aligned.
5. It handles the entire calibration process, including data extraction from the Growth Research Database (GRD), data cleaning, and calibration. If the existing tool needs to undergo a third-party review, which is not uncommon when bidding for a large contract, it is nearly impossible to provide that.

While an alternative is to enhance the existing item calibration system, it is more desirable to build a stand-alone tool that uses a programming language familiar to psychometricians and allows them to exercise autonomy in its use. Specifically, a software application is needed to allow psychometricians to choose analysis and options, to allow operational analysis or QA of data from any intermediate procedure, to better protect data security, and to allow the software to be shared with third parties for independent replication. As such, Pychometrik was developed. It was built in Python and has introduced some changes to the item calibration process, including adopting a curve-fitting algorithm to estimate item difficulty. Aside from being equipped with features to support the item calibration process updates described above, the following procedures from the existing process were excluded from Pychometrik: (1) iterative grade range (IGR) procedure, (2) two-pass filtering, and (3) Model of Man (MoM) procedure.

2. Item Calibration Process

2.1. Existing Tool

Built with Microsoft’s T-SQL language, the existing item calibration tool can handle the entire item calibration process from within the GRD. Calibration is manually initiated and automatically runs a series of analyses, including steps for data harvesting and cleaning, item calibration, and post-calibration analysis and synthesis. This automation allows for continuous field test item calibration and the release of resources and time for other tasks due to the automated decision-making on the item calibration status. The current item calibration process operates in the following major steps, as illustrated in Figure 2.1.

Step 1: Data cleaning. Apply a series of pre-determined rules to filter out invalid test events and responses and identify what test events and responses to use in calibration.

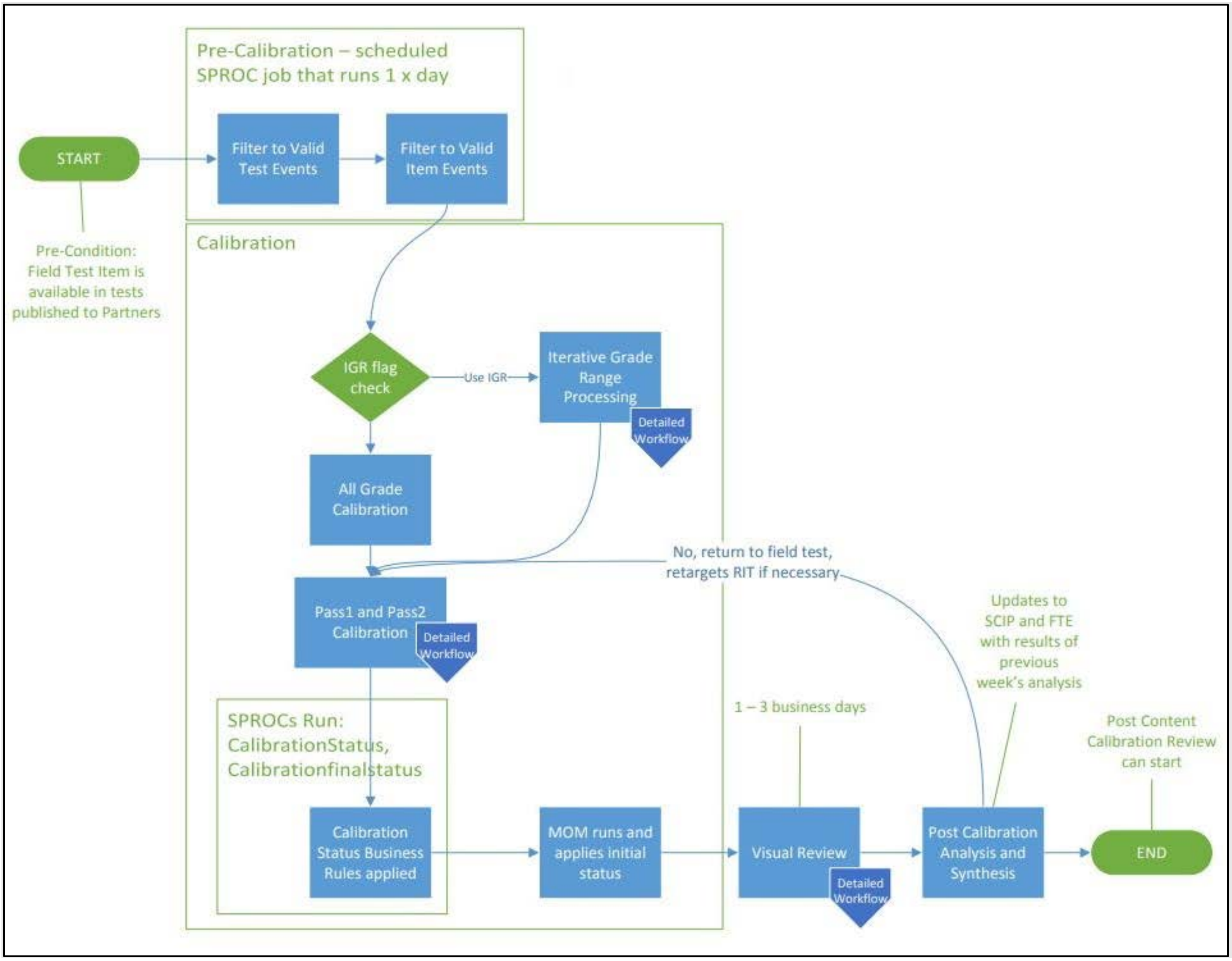
Step 2: Item parameter estimation. Derive item difficulty estimates from a calibration sample determined by either the all grade calibration (AGC) or iterative grade range (IGR) procedure. Regardless of which calibration sample is used, a maximum likelihood estimation (MLE) type procedure by fixing person ability is used to estimate item difficulty (i.e., RIT), along with two-pass filtering.

a. AGC/IGR. AGC identifies a calibration sample consisting of all students exposed to the same item from Step 1, whereas IGR uses an iterative procedure to identify the best fitting grade(s) exposed to the same item from Step 1 and uses that subset of student responses to derive item parameter estimates. Determining whether to use AGC or IGR is evaluated by a chi-square goodness-of-fit test using the lowest grade with at least 25 responses. If the chi-square test is passed, AGC will be used. If not, IGR is used.

b. Two-pass filtering. Two-pass filtering results in two item parameter estimates for an item regardless of the use of AGC or IGR. Pass 1 calibration is derived from responses from a full calibration sample identified by either AGC or IGR. Pass 2 calibration is derived from responses excluding test events from the calibration sample with less than 10% probability of answering the item correctly, given Pass 1 item calibration. An “official” item parameter estimate is determined from Pass 2. Studies indicate that item parameter estimates from Pass 1 and Pass 2 are generally comparable with each other (He, 2018; Meyer & Bo, 2019). For example, suppose an item was calibrated with 210 RIT (i.e., 1 logit) from Pass 1. Using the Rasch model, *p = 1/1+exp (b-theta)* with *p = 0.1* and *b = 1.0* to get theta, which is 188 RIT. This theta is used as a cutoff to remove test events with a final ability estimate lower than theta. Test events with a final ability estimate smaller than 188 are removed, and the remaining test events are used to obtain Pass 2 calibration. This procedure results in a field testing inefficiency because very large numbers of students are excluded from calibration of very difficult items, and items remain in field testing for a long time.

Step 3: Model of Man (MoM). The MoM procedure (Hauser et al., 2014) generates an automated item calibration status for each item to identify items that need further visual inspection by psychometricians. It uses a logistic modeling approach to create a model that mimics how item reviewers integrate item-level fit statistics and graphical performance plots to predict the item reviewer's assignment of the item's calibration status. The MoM procedure was implemented in the calibration system to reduce the number of items that need human review and monitor the consistency of item review decisions within and across different human reviewers.

Figure 2.1. Existing Item Calibration Workflow


Pre-Calibration – scheduled SPROC job that runs 1 x day
START
Filter to Valid Test Events
Filter to Valid Item Events
Pre-Condition: Field Test Item is available in tests published to Partners
Calibration
IGR flag check
Use IGR
Iterative Grade Range Processing
Detailed Workflow
All Grade Calibration
No, return to field test, retargets RIT if necessary
Pass1 and Pass2 Calibration
Detailed Workflow
Updates to SCIP and FTE with results of previous week's analysis
SPROCs Run: CalibrationStatus, Calibrationfinalstatus
1 – 3 business days
Post Content Calibration Review can start
Calibration Status Business Rules applied
MOM runs and applies initial status
Visual Review
Detailed Workflow
Post Calibration Analysis and Synthesis
END

Item difficulty estimation is conducted in the existing tool with a brute force approach that fixes person ability estimates to values estimated from operational items. By fixing the ability estimates of the students administered an individual field test item, the algorithm searches for a RIT value ranging between 100 and 350 that can minimize the mean square fit (MSF) and treats that value as item difficulty on the RIT scale of interest.

Table 2.1 and the following equations describe how to calculate the MSF for each brute-force grid given item responses and student ability estimates for an individual item that needs to be calibrated. In general, students who have been administered an individual item are grouped together based on their ability estimates at an interval of 1 RIT. For each *i*th ($i = 1,2,3,4,\dots,n$) ability level, MSF conditional on each *j*th ($j = 1,2,3,4,\dots,m$) RIT grid between 101 and 350 is calculated based on Equation 1. Summing up the MSFs across all ability levels, b_j that can minimize the MSF becomes item difficulty for an individual item. Both Pass 1 and Pass 2 procedures use the same algorithm to calculate item difficulty, except that Pass 2 removes a proportion of students from the calibration sample whose probabilities of answering the item correctly are less than 10% given the item parameter estimates obtained in Pass 1. The item parameter estimate from Pass 2 is used as the "official" item difficulty for an individual item.

Table 2.1. Mean Square Fit (MSF) Calculation

				b (Item Difficulty)					
				b_1(101)	b_2(102)	b_3(103)	b_4(104)	…..	b_m(350)
				b_1(-9.9)	b_2(-9.8)	b_3(-9.7)	b_4(-9.6)		b_m(15)
RIT (Student Final Ability)	Logit (Student Final Ability)	w (Low Count Weight)	n (Observed Correct Score)	$N_{P_{b_1}}$ (Expected Correct Score)	$N_{P_{b_2}}$ (Expected Correct Score)	$N_{P_{b_3}}$ (Expected Correct Score)	$N_{P_{b_4}}$ (Expected Correct Score)	…..	$N_{P_{b_m}}$ (Expected Correct Score)
RIT_1	Θ_1	w_1	n_1	$N_{11}*p_{1b_1}$	$N_{12}*p_{1b_2}$	$N_{13}*p_{1b_3}$	$N_{14}*p_{1b_4}$	…..	$N_{1m}*p_{1b_m}$
RIT_2	Θ_2	w_2	n_2	$N_{21}*p_{2b_1}$	$N_{22}*p_{2b_2}$	$N_{23}*p_{2b_3}$	$N_{24}*p_{2b_4}$	…..	$N_{2m}*p_{2b_m}$
RIT_3	Θ_3	w_3	n_3	$N_{31}*p_{3b_1}$	$N_{32}*p_{3b_2}$	$N_{33}*p_{3b_3}$	$N_{34}*p_{3b_4}$	…..	$N_{3m}*p_{3b_m}$
…..	…..	…..	…..	…..	…..	…..	…..	…..	
RIT_n	Θ_n	w_n	n_n	$N_{n1}*p_{nb_1}$	$N_{n2}*p_{nb_2}$	$N_{n3}*p_{nb_3}$	$N_{n4}*p_{nb_4}$	…..	$N_{nm}*p_{nb_m}$
				MSF_{b_1}	MSF_{b_2}	MSF_{b_3}	MSF_{b_4}	…..	MSF_{b_m}

$$MSF_{b_j} = \frac{\sum_{i=1}^{n} w_i * \left(n_i - N_{ij} * p_{ib_j}\right)^2}{\sum_{i=1}^{n} w_i * n_i} \tag{1}$$

$$p_{ib_j} = \frac{\exp\,(\theta_i - b_j)}{1 + \exp\,(\theta_i - b_j)} \tag{2}$$

where

$$RIT_i = Student\ Final\ Ability\ Estimate \quad i = 1,2,3,4,\dots,n$$

$$\theta_i = \frac{RIT_i - 200}{10}$$

$$w_i = \begin{cases} 1\ if\ count \geq 5 \\ 0.01\ if\ count < 5 \end{cases}$$

$$b_j\!: Every\ Possible\ Item\ RIT \quad j = 1,2,3,4,\dots,m$$

2.2. Pychometrik

Pychometrik consists of a series of classes and functions written in Python that can conduct both classical item analysis and item calibration for the Rasch family models. It is designed to be extendable and incorporate additional IRT models such as the two-, three-, and four-parameter models, as well as other types of psychometric analysis. For Rasch item calibration, a proportional curve fitting (PCF) algorithm described in Meyer and Hailey (2012) is used to estimate item parameters. Fixed-person item calibration is achieved by fixing person ability estimates to values obtained from operational items and then estimating item difficulty to bring new items onto the same scale as the old items (i.e., item measures are estimated by anchoring persons at pre-set, fixed measures). This PCF algorithm is also used in WINSTEPS and jMetrik. Both the IGR and Pass 2 procedures from the existing calibration process are not included in Pychometrik.

The PCF algorithm implemented in Pychometrik uses a logistic function and a function that describes the regression of a score on a measure. Parameter estimates are on the logit scale. For $i = 1, \dots, n$ ability levels, there are w_i examinees scoring at level i with an ability value of θ_i. The estimation process works in an iterative manner until the convergence criterion is met, as described below.

1. Compute the expected composite item scores based on the known ability values and current difficulty estimate, *b*, using Equation 3.

$$T_b = \sum_{i=1}^{n} w_i \frac{\exp\,(\theta_i - b)}{1 + \exp\,(\theta_i - b)} \tag{3}$$

2. Also compute the expected composite item score with the current difficulty value plus a small amount, $b + \Delta$, using Equation 4. At the start of the estimation process, b is set to zero and Δ is set to 1.

$$T_{b+\Delta} = \sum_{i=1}^{n} w_i \frac{\exp\,(\theta_i - [b+\Delta])}{1 + \exp\,(\theta_i - [b+\Delta])} \tag{4}$$

3. Compute the observed composite item score, S_+, which is the sum of the item score for all students responding to the item. When students are grouped into a two-way table, according to the $i = 1, \dots, n$ ability levels (rows) and the $k = 1, \dots v$ possible item score values, u_k, (columns), each cell is a frequency count, w_{ik}, and the observed item score is computed by Equation 5.

$$S_+ = \sum_{i=1}^{n} \sum_{k=1}^{v} w_{ik} u_k \tag{5}$$

4. Convert an item score to the logistic ogive using Equation 6. The item score is generically denoted as x. It may be either the expected composite item score from Equation 3 or Equation 6, or the observed composite item score.

$$f(x) = log \left\{ \frac{x - min_x}{max_x - x} \right\} \tag{6}$$

   where min_x and max_x represent the smallest and largest possible values of x, respectively, and log indicates the natural logarithm.

5. The updated item difficulty estimate, b^*, is the value that moves the expected logistic curve closer to the observed logistic curve using Equation 7.

$$b^* = slope * f(S_+) + intercept \tag{7}$$

where

$$slope = \frac{\Delta}{f(\mathrm{T}_{\mathrm{b}+\Delta}) - f(T_b)} \tag{8}$$

$$Intercept = b - slope * f(T_b) \tag{9}$$

Combining Equations 7–9 shows that the updated item difficulty estimate is given by:

$$b^* = b + \left\{\frac{\Delta[f(S_+) - f(T_b)]}{f(T_{b+\Delta}) - f(T_b)}\right\}, \tag{10}$$

where the difficulty estimate is changed proportional to the difference of the observed and expected composite item scores.

To prevent a drastic change in the parameter estimate at any iteration, do not change the estimate by more than one logit using Equation 11.

$$b^* = \max\left(\min(b + 1,\ b^*),\ b - 1\right) \tag{11}$$

6. Set $\Delta = |b^* - b|$.

7. Repeat Steps 1–6 using b^* as the most current item difficulty estimate until the Δ is less than or equal to the convergence criterion.

3. Results

3.1. Item Parameter Estimate Comparability Study

The goal of the item parameter estimate study was to evaluate the comparability of the item calibration results from the two systems, including item parameter estimates, percent correct (Pvalue), infit, outfit, point bimeasure (Pbm), and correlation between the empirical and theoretical probability of correctness (ExpCorr). Both a real-data study and a simulation study were conducted.

3.1.1. Real Data

The real-data study began with a random sample of 1,635 MAP Growth items that were successfully calibrated after January 1, 2018, as shown in Table 3.1. Each item's calibration sample included students who were administered the item after it became operational. Specifically, item responses and students' final ability estimates were extracted from the GRD, and item calibration was conducted in both tools by anchoring on the students' final ability estimates. However, while the study intended to include 1,635 items, the existing tool returned the calibration results for only 1,591 items. Among these 1,591 items, 1,180 items were calibrated via the AGC procedure and 411 via IGR. Pychometrik returned the calibration results for all 1,635 items.

Table 3.1. Number of Items in the Item Parameter Estimate Comparability Study using Real Data

Scale	#Items	%
Mathematics	790	48.32
Reading	278	17.00
Language	128	7.83
Science	232	14.19
Spanish Reading	207	12.66
Total	1,635	100.00

The item calibration results from the two tools were compared in terms of item parameter estimates, Pvalue, infit, outfit, Pbm, and ExpCorr. To make the results from both tools comparable, results from the existing tool used for comparison were those from Pass 1, and item parameter estimates from Pychometrik were rounded to one decimal point in logit and the relevant item statistical indices such as infit and outfit were computed using the rounded item parameter estimates. Since Pychometrik removed IGR from the process, only the 1,180 items calibrated via AGC in the existing tool were used for the comparison. Among these 1,180 items, both tools returned identical calibration results for 663 items. The remaining 517 items had different calibration results, as shown in Table 3.2 that presents the overall results. The calibration results from the existing tool served as the basis for comparison.

As shown in Table 3.2, the average item RITs from the existing tool are slightly larger than those from Pychometrik across all five scales, with differences ranging from 0.33 RIT for mathematics to 0.91 RIT for science. The differences in item RITs seem to have affected both item infit and outfit more than Pvalue, Pbm, and ExpCorr, for which the average differences are almost zero if rounded to the second decimal point for ExpCorr and to the third decimal point for both Pvalue and Pbm. The average differences in infit and outfit suggest that item parameter estimates from Pychometrik provide better item fit than those from the existing tool.

Table 3.2. Overall Item Calibration Results and Differences for the 517 Items

		Item RIT		Difference					
Scale	N	Existing	Pychometrik	RIT	Pvalue	Pbm	Infit	Outfit	ExpCorr
Mathematics	234	227.84	227.51	-0.33	0.00	0.00	-0.03	-0.12	-0.00
Reading	78	201.10	200.35	-0.76	0.00	0.00	-0.01	-0.02	-0.00
Language	41	209.66	208.78	-0.88	0.00	0.00	-0.03	-0.10	-0.00
Science	80	223.16	222.25	-0.91	0.00	0.00	-0.03	-0.08	-0.00
Spanish Reading	84	192.58	191.81	-0.77	0.00	0.00	-0.02	-0.06	-0.00
Overall	517	215.91	215.31	-0.60	0.00	0.00	-0.03	-0.09	-0.00

Table 3.3 presents the number of items by RIT difference. The item parameter estimates for most items had a 1 RIT difference between the two tools (318 + 137 = 455 items, or 88%). This was expected due to rounding. Pychometrik does not round estimates during the iterative procedure. It only rounds the final result. By contrast, the brute-force procedure used by the existing tool uses only one decimal point, as Table 2.1 indicates, so it does not allow for more precision than one decimal.

Table 3.3. Distribution of Items by RIT Differences for the 517 Items

RIT Difference	Overall		#Items				
	N	%	Math	Reading	Language	Science	Spanish Reading
-7	1	0.19	1	–	–	–	–
-5	4	0.77	1	–	–	–	3
-4	3	0.58	1	–	–	1	1
-3	5	0.97	2	–	1	–	2
-2	44	8.51	22	4	7	4	7
-1	318	61.51	110	63	26	68	51
1	137	26.50	96	10	7	7	17
2	2	0.39	–	1	–	–	1
3	2	0.39	–	–	–	–	2
4	1	0.19	1	–	–	–	–
Total	517	100.00	234	78	41	80	84

The following steps were taken to understand why some item RIT differences between the tools were larger than 2:

1. Independently develop an estimation code that mimics the item calibration procedures used by both tools and run item calibration for the 517 items. The item parameter estimates output from both tools were successfully replicated by the independently created code.
2. Use a Newton-Raphson procedure to calibrate these items. Only the item parameter estimates from Pychometrik were replicated.
3. Compare the log-likelihood based on item parameter estimates from the two systems. A likelihood function measures the goodness-of-fit of a statistical model to a sample of data for given values of the unknown parameters. A higher log-likelihood value suggests better model-data fit than a lower log-likelihood value.

Table 3.4 presents the average log-likelihood values for all 1,180 AGC items based on the item RITs and the 517 items with RIT differences from the two systems. The results suggest that, on average, Pychometrik provides better item parameter estimates than the existing tool.

Table 3.4. Descriptive Statistics of Log-likelihood

		Log-likelihood Descriptive Statistics			
Tool	**#Items**	**Mean**	**SD**	**Min.**	**Max.**
1,180 AGC Items					
Existing	1,180	-2803.94	3685.89	-85083.56	-81.37
Pychometrik	1,180	-2801.66	3684.17	-85083.56	-81.37
Difference	1,180	2.28	12.24	0.00	373.26
517 Items with RIT Differences					
Existing	517	-2739.59	2563.32	-28330.55	-85.25
Pychometrik	517	-2734.39	2557.52	-28288.70	-83.75
Difference	517	5.20	18.09	0.00	373.26

3.1.2. Simulations

For the simulation study, 60 Rasch items were randomly generated out of the standard normal distribution and administered to a group of 1,200 simulees with abilities distributed around the standard normal distribution. Both tools were used to calibrate these items by using the fixed-person calibration design (i.e., fixing the students' abilities). The results were compared in two ways: (1) compare the item parameter estimates produced by each tool and (2) compare the item parameter estimates with true item parameters. The bias, mean absolute deviation (MAD), and mean square error (MSE) indices defined below were used to evaluate the results in each comparison, where β_i is true item difficulty for item i and $\hat{\beta}_i$ is difficulty estimate for item i:

$$Bias = \frac{\sum_{i=1}^{60}(\hat{\beta}_i - \beta_i)}{60}$$

$$MAD = \frac{\sum_{i=1}^{60}(|\hat{\beta}_i - \beta_i|)}{60}$$

$$MSE = \sqrt{\frac{\sum_{i=1}^{60}(\hat{\beta}_i - \beta_i)^2}{60}}$$

Table 3.5 presents the summary statistics of the item parameter estimates from both tools, including how the estimates compare with each other and with the true item parameters. In general, the item parameter estimates from both tools are comparable to the true item parameters, but those from Pychometrik are slightly closer to the true parameters than those from the existing tool with a smaller standard deviation (i.e., 9.47 (Pychometrik) vs. 9.46 (true) vs. 9.51 (existing tool)). The bias results also indicate that, on average, the item parameter estimates from the existing tool are underestimated by 0.03 RIT compared with the true item parameter estimates.

Table 3.5. Summary Statistics of RIT Item Parameter Estimates in the Simulation Study

Statistics	#Items	Mean	SD	Min.	Max.
RIT					
Existing	60	199.78	9.51	176	220
Pychometrik	60	199.82	9.47	177	220
True	60	199.82	9.46	177	220
Pychometrik vs. True					
Bias	60	0.00	0.32	-1	1
MAD	60	0.10	0.30	0	1
MSE	60	0.10	0.30	0	1
Existing vs. True					
Bias	60	-0.03	0.32	-1	1
MAD	60	0.10	0.30	0	1
MSE	60	0.10	0.30	0	1
Pychometrik vs. Existing					
Bias	60	0.03	0.32	-1	1
MAD	60	0.10	0.30	0	1
MSE	60	0.10	0.30	0	1

3.2. Pychometrik Item Parameter Recovery Simulation Study

3.2.1. Design and Analysis

The goal of this study was to test the Pychometrik software by examining the item parameter recovery of the new tool using the on-grade MAP Growth student distributions and item parameter distributions by running simulations. The study was conducted using MAP Growth Reading score distributions in nine grades from Grades K–8 with three sample sizes (500, 1,000, and 2,000) for a total of 27 conditions. One thousand item difficulty values (on the RIT scale) were generated from a uniform distribution ranging from 100 to 350, and converted to the logit scale using the formula below:

$$b_{logit} = \frac{RIT - 200}{10}$$

Table 3.6 presents the descriptive statistics of the 1,000 simulated item difficulties (on the logit scale) across all grades in MAP Growth Reading. The mean of the item difficulty is 2.37 and the values range from -9.95 to 14.98. Table 3.7 presents the number of items in each of the 10 bins that group the items in various item difficulty ranges. Each bin has 100 items on average. The generating uniform distribution has a mean of 2.5, a minimum of -10, and a maximum of 15 when converted to logits.

Table 3.6. Descriptive Statistics of the Simulated Item RITs

		Item Difficulty Descriptive Statistics (in Logits)						
						Percentile		
Scale	N	Mean	SD	Min.	Max.	25%	50%	75%
Reading	1,000	2.37	7.24	-9.95	14.98	-3.95	2.22	8.45

Table 3.7. Number of Simulated Items in a Bin

Bin	#Items
(-10.0, -7.5]	97
(-7.5, -5.0]	114
(-5.0, -2.5]	99
(-2.5, 0.0]	98
(0.0, 2.5]	110
(2.5, 5.0]	104
(5.0, 7.5]	86
(7.5, 10.0]	90
(10.0, 12.5]	98
(12.5, 15.0]	104

Simulees were generated for each grade based on the corresponding on-grade MAP Growth Reading RIT score distribution based on real data from the MAP Growth technical report, as shown in Table 3.8. The last two columns are the corresponding values on the logit scale.

Table 3.8. RIT Distributions

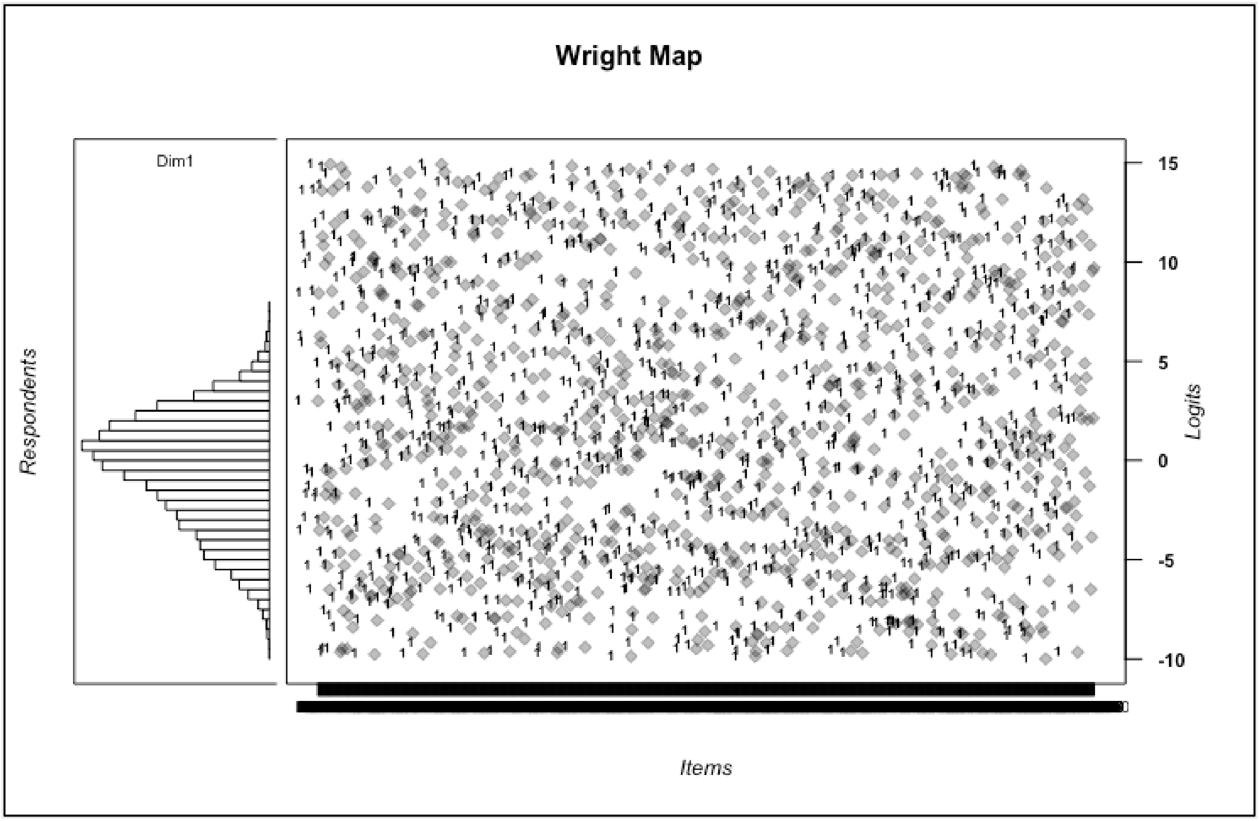
	RIT		Logits	
Grade	Mean	SD	Mean	SD
K	150.10	14.40	-4.99	1.44
1	169.40	16.70	-3.06	1.67
2	182.20	17.30	-1.78	1.73
3	193.20	17.00	-0.68	1.70
4	201.90	16.30	0.19	1.63
5	208.40	16.00	0.84	1.60
6	212.90	15.90	1.29	1.59
7	216.70	16.00	1.67	1.60
8	220.30	16.20	2.03	1.62

Figure 3.1 displays the Wright map (Boone & Scantlebury, 2006; Wilson, 2004) to visually display the simultaneous distributions of items and simulees. The distribution of simulees (on the left) and items (on the right) are displayed on the same logit scale. Table 3.9 presents the overall descriptive statistics of the two distributions. The mean simulee ability is centered around 0, and the mean item difficulty is centered around 2. The minimum values are comparable, but the maximum value of simulee ability is much smaller than the maximum value of item difficulty.

Table 3.9. Overall Descriptive Statistics of Simulee Ability and Item Difficulty

		Descriptive Statistics				
Distribution	N	Mean	SD	Median	Min.	Max.
Simulee Ability	18,000	-0.48	2.75	-0.10	-9.94	7.94
Item Difficulty	1,000	2.37	7.24	2.22	-9.95	14.98

Figure 3.1. Wright Map of Simulee Ability and Item Difficulty



To generate the item responses, the following procedures were used:

- Generate a probability matrix based on the simulee and item parameters (in logits) according to the Rasch model.
- Draw a matrix of random numbers from a uniform distribution with a minimum of 0 and a maximum of 1 (U[0,1]).
- For each cell in the matrix, compare the random number to the probability of a correct response to create a dichotomous item response. If the random number is greater than the probability, the response is a 0, while a probability greater than the random number results in a response of 1.

To assess the item parameter recovery, the following indexes were calculated:

- Bias = $\frac{\sum_{i=1}^{100}(\hat{\beta}_i - \beta_i)}{100}$, where β_i is a given item parameter, and $\hat{\beta}_i$ is its estimate.
- Mean absolute deviation (MAD) = $\frac{\sum_{i=1}^{100}(|\hat{\beta}_i - \beta_i|)}{100}$
- Root mean square error (RMSE) = $\sqrt{\frac{\sum_{i=1}^{100}(\hat{\beta}_i - \beta_i)^2}{100}}$
- Correlation between the simulated item parameters and the estimated item parameters for each condition

3.2.2. Results

Figure 3.2, Figure 3.3, and Figure 3.4 display the RMSE, MAD, and correlation results, respectively, from the item parameter recovery simulation study. The y-axis is the index, and the x-axis is the grade. The three different colors represent the three different sample sizes. Item recovery for Grade 8 students has the best accuracy in all three figures. The Grade 8 score distribution has a mean of 2.03 (see Table 3.7), which is the closest to the mean item difficulty of 2.37 (see Table 3.8) compared to the other grades. Another observation is that the higher the sample size, the higher the accuracy of item parameter recovery.

Table 3.10 summarizes the results when the range of item difficulties is the same as the range of the simulee abilities with the minimum value being -10 and the maximum value being 8. Both MAD and RMSE are small across all three sample sizes, and the correlations are all above 0.99. These results indicate that Pychometrik is working properly.

Table 3.10. Overall Item Recovery Results

Sample Size	MAD	RMSE	Correlation
500	0.45	0.75	0.99
1,000	0.34	0.59	0.99
2,000	0.27	0.47	1.00

Figure 3.2. RMSE between Estimated and Simulated Item Parameters

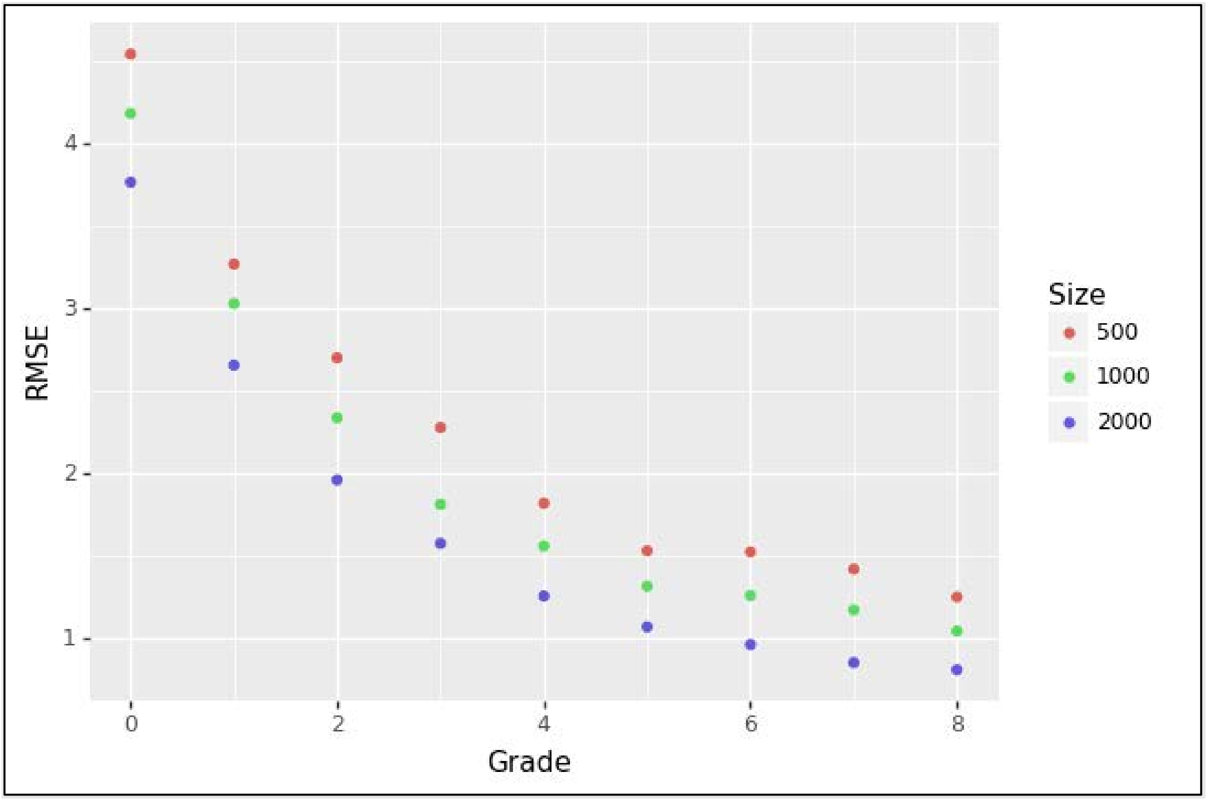


Figure 3.3. MAD between Estimated and Simulated Item Parameters

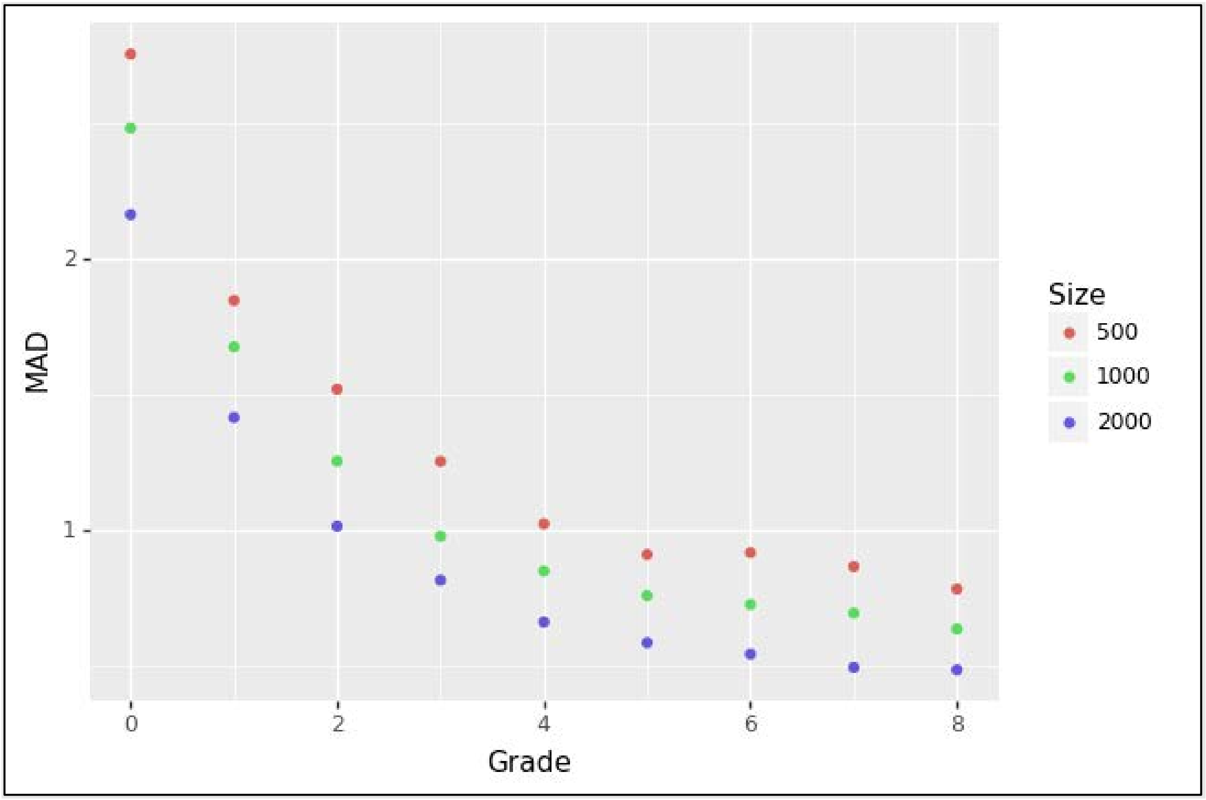


Figure 3.4. Correlation between Estimated and Simulated Item Parameters

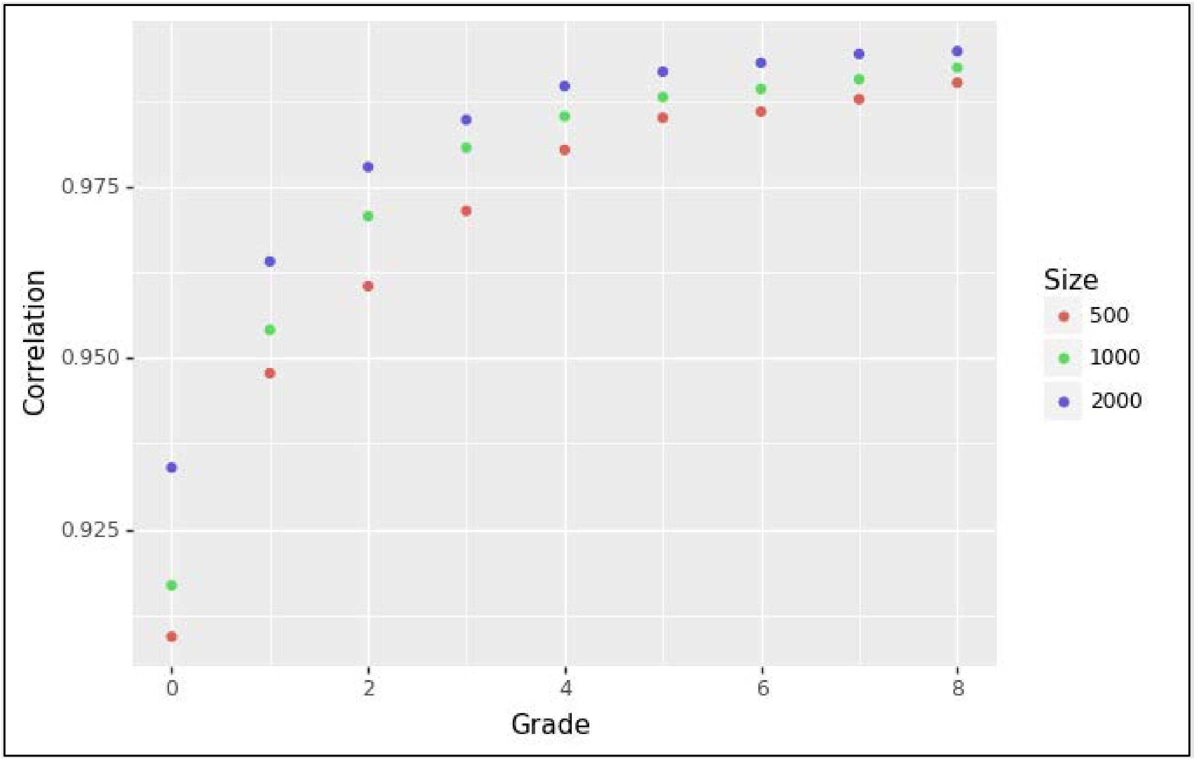


Figure 3.5, Figure 3.6, and Figure 3.7 present the results by item parameter bins. As recalled in the Wright map above, there are no simulees with an ability above 7.94. Thus, the results show high RMSE, MAD and low correlations above that point on the right end of the x-axis in all three figures. Item recovery using the Grade 8 simulees has the best results among all grades because the distribution is simulees in that grade is similar to the distribution of items.

Figure 3.5. RMSE by Item Parameter Bin (Sample Size of 2,000)

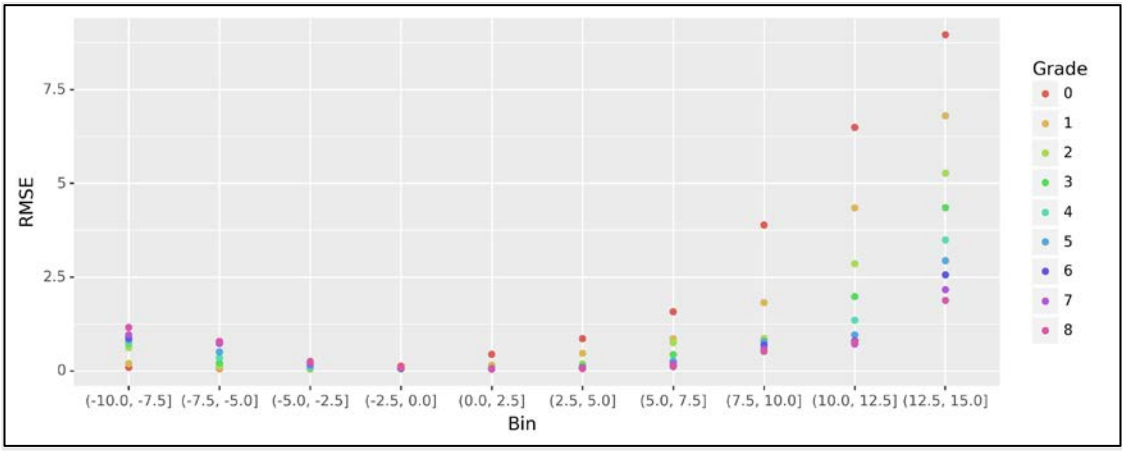


Figure 3.6. MAD by Item Parameter Bin (Sample Size of 2,000)

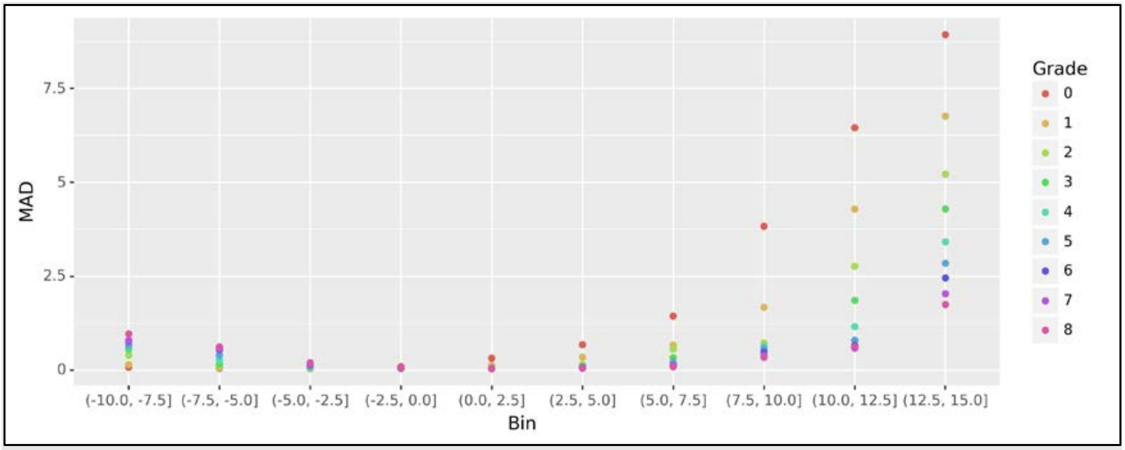
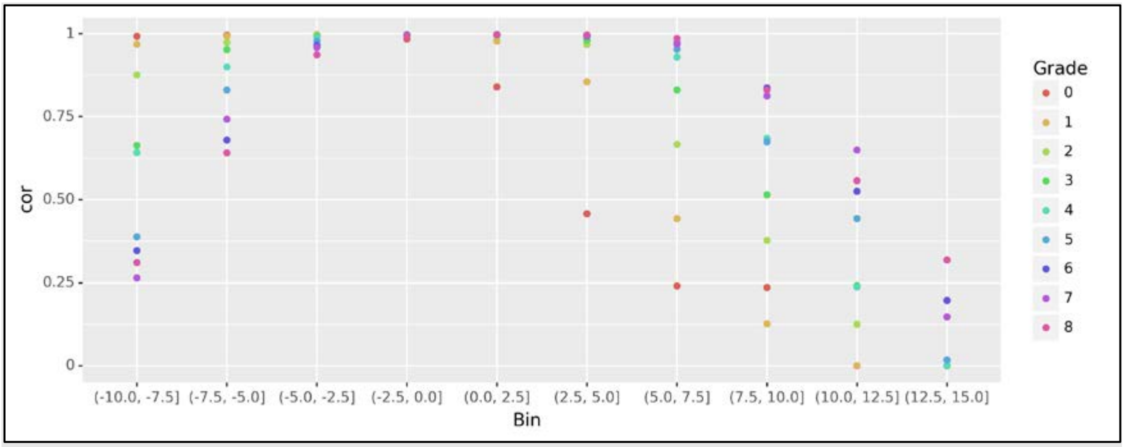


Figure 3.7. Correlation by Item Parameter Bin (Sample Size of 2,000)



4. Summary and Conclusion

Overall, the results from both studies presented in this report support the use of Pychometrik for item calibration. Results from the item parameter estimate comparability study suggests that item calibration results from both tools are mostly comparable with each other, including item parameter estimate, Pvalue, infit, outfit, Pbm, and ExpCorr. The analyses conducted to understand why some item RIT differences between the two tools were larger than 2 suggest that, in some cases, Pychometrik can produce more accurate results. The Pychometrik item parameter recovery simulation results show that the estimates of the item parameters are accurate when both the ranges of item parameters and the simulee parameters align, supporting the use of the new tool for item calibration.

5. References

Boone, W. J., & Scantlebury, K. (2006). The role of Rasch analysis when conducting science education research utilizing multiple-choice tests. *Science Education, 90*(2), 253–269.

Hauser, C., Thum, Y., He, W., & Ma, L. (2014). Using a model of analysts' judgments to augment an item calibration process. *Educational and Psychological Measurement, 75*(5), 826–849.

He, W. (2018). *Differences in item calibrations from Pass I and Pass II procedures*. NWEA.

Meyer, P., & Bo, E. (2019). *Exploring filtering responses in a two-pass calibration process*. NWEA.

Meyer, P., & Hailey, E. (2012). A study of Rasch, partial credit, and rating scale model parameter recovery in WINSTEPS and jMetrik. *Journal of Applied Measurement, 13*(3), 248–258.

Wilson, M. R. (2004). *Constructing measures: An item response modeling approach*. Lawrence Erlbaum Associates.