Calibration of Spanish MAP Growth Math Tests

June 2020

Shudong Wang, NWEA Psychometric Solutions
Sylvia Li, NWEA Psychometric Solutions

nwea

Suggested citation: Wang, S., & Li, S. (2020). *Calibration of Spanish MAP Growth math tests*. Portland, OR: NWEA.

Table of Contents

List of Tables

1. Introduction

This document presents the results of a calibration study conducted to investigate the possible consequences of replacing the English MAP® Growth™ Math item parameters with the Spanish math item parameters from calibrating Spanish items using empirical data. Spanish math items currently have the same item parameters as their English counterparts, but calibrating the Spanish instead using empirical student data will result in a more reliable and valid assessment. The final calibration results are based on monolingual data only (i.e., Spanish test responses), but bilingual results are also used as references.

1.1. Background

To equate the original Spanish MAP Growth Math scale to the English scale, NWEA used Rasch calibration and a common-item equating method. The common items used in the equating were nonverbal items (i.e., items with very few words such as 1 + 1 = 2) or items with limited association with linguistic content. The original Spanish math scale was equated based on only a few hundred items in the Spanish math item pool that were transadapted from the English MAP Growth Math items. After expanding the Spanish math item pools, NWEA used back translation design (BTD) to double-check the transadaptation work. BTD requires translating the source version of the test (English) into the target language (Spanish), then translating them back to English and comparing them with the source language to identify possible discrepancies. After any reconciliation with the content team, the English math item parameters could be used as the Spanish item parameters.

Since equating the Spanish math items to the English math scale based on a small number of items, NWEA has not calibrated any Spanish items using student responses. Instead, NWEA created the Spanish math tests by using English item parameters. However, this practice has the following two major issues. This report only focuses on the first issue.

1. Appropriateness of using English item Rasch Units (RITs) as Spanish item RITs instead of empirical data
2. Removal of any discontinuity of scale between the Spanish MAP Growth Math K–2 and 2–5 tests

About five years ago, NWEA decided to use the English item RIT scores for the Spanish math items because it was impossible at the time to collect enough student responses to calibrate the Spanish items. In other words, all Spanish math items have been transadapted from English items using their English RIT counterpart because the Spanish items are not calibrated. Spanish scores are therefore equivalent to the English item RITs. According to both classical test theory and item response theory (IRT), the scale that a test is designed to create and the construct that a test intends to measure can only be made and verified based on student's empirical responses. When enough student responses are collected, the Spanish math item parameters should be calibrated based on empirical data and the Spanish math scale based on item parameters should be re-established and validated. Because NWEA uses the Rasch model to create and equate all MAP Growth scales, the calibration of items using student responses serves two purposes: (1) create a new scale and (2) equate new items to an existing scale. Therefore, it is preferable to calibrate Spanish item RIT based on empirical data to ensure an accurate representation of student performance

1.2. Research Questions

There is a need to better understand the impact of using English item RITs for transadapted Spanish items. The following research questions are therefore addressed in this calibration study:

1. What are the differences in item RITs for the Spanish transadapted items between using English item RITs and the calibrated Spanish item RITs when fixing Spanish student RITs scored using English item RITs for both monolingual and bilingual students?

2. What are the differences in student RIT scores for Spanish transadapted items between using English item RITs and calibrated Spanish item RITs that include item calibration status 10 items when fixing Spanish student RITs scored using English item RITs for monolingual students?

3. What are the differences in student RIT scores for Spanish transadapted items between using English item RITs and calibrated Spanish item RITs that do not include item calibration status 10 items when fixing Spanish student RITs scored using English item RITs for monolingual students?

4. What are the differences in student RIT scores for Spanish transadapted items between using English item RITs and calibrated Spanish item RITs that include item calibration status 10 items when fixing Spanish student RITs scored using English item RITs for bilingual students?

2. Method

2.1. Calibration Designs

Students take two tests, one in Spanish and one in English. The English items are used to score the English math test. This English score is then used to calibrate the Spanish math items. To check the impact of different calibration procedures to determine which one is most effective, three calibration designs for Spanish item and person parameters employed the fixed-person score calibration method in which the old Spanish person RITs were fixed while calibrating the new Spanish item parameters, as summarized in Table 2.1. The item RIT refers to each item's RIT designation, whereas person RIT refers to student scores. Old RITs refer to the previous RITs obtained using the English item parameters, whereas the new RITs are being obtained with this calibration. After calibrating the new Spanish item parameters, the new parameters were used to obtain the new Spanish person parameters. In this way, the newly calibrated Spanish item and person parameters are equated to the English math scale. In all calibrations, the old transadapted Spanish item parameters are equal to the values of their counterpart English item parameters (i.e., English item RITs = old Spanish item RITs). Once an item has been calibrated, it is labeled with one of the following calibration statuses:

1. Calibration status 10 = items that need to be re-field tested and re-calibrated to accumulate more responses in the item calibration procedure
2. Calibration status XX = field tested items that pass the calibration
3. Calibration status 7, 12, or 13 = field tested items that are rejected during calibration and undergo content review

Based on these statuses, the following scoring procedures for the new Spanish person RITs were implemented to determine the impact of using status 10 items in scoring. If there is no impact when comparing the two different scoring procedures, status 10 items could be used in an operational test.

1. Score students using the new Spanish item RITs with calibration status 10 and calibration status XX.
2. Score students using the new Spanish item RITs with calibration status XX only.

Distinguishing the utility of item status 10 in calibration is meant to check the impact of these items on newly calibrated Spanish item and person parameters. During the calibration, items are labeled as status 10 when that calibration sample size is less than 1,000. The monolingual data are from all students who have Spanish math test results only, and bilingual data are from all students who have both Spanish and English math test results. The monolingual data include students who took Spanish math tests in the bilingual data.

Table 2.1. Spanish Item Calibration Designs

Design	Student Data	English	Spanish					Fixed in Calibration
		Item RIT	Old Item RIT	New Item RIT	Old Person RIT	New Person RIT	Response	
1	Monolingual	Used	Used	Created	Used	Created with Item Status 10	Used	Old Spanish Person RIT
2	Monolingual	Used	Used	Created	Used	Created without Item Status 10	Used	Old Spanish Person RIT
3	Bilingual	Used	Used	Created	Used	Created with Item Status 10	Used	Old Spanish Person RIT

2.2. Data

2.2.1. Item Data

Based on the item data pulled from the NWEA Growth Research Database (GRD) system for the past five years, NWEA has accumulated 3,908 Spanish math items across different tests. In operation, the same item can be used in different tests. The two types of item samples used in this report are collected and selected samples. The collected sample is the originally pulled sample, and the selected sample is the sample of items filtered using the following item selection criterion:

> *Absolute difference between old Spanish item RIT (equivalent to the English item RIT) and new Spanish item RIT (obtained from calibration using Spanish item responses) is less than or equal to 20 RITs*

Table 2.2 and Table 2.3 present the number of items by test and item calibration status for both the collected and selected samples with duplication (i.e., data are duplicated because the same items can be used in different tests). The original item calibration status (obtained from the English test calibrations) for all Spanish item calibration statuses (i.e., 07, 10, 12,13, XX) is XX. As shown in these tables, the collected sample had a total of 3,955 items, and the selected sample had a total of 3, 908 items. This indicates that 47 items have differences between the old and new RITs larger than 20 RITs, which is about 1.1 % of total number of items in the system. These items are regarded as outliers from calibration process.

As shown in Table 2.3, these outlier items come from different calibration statuses. All items with calibration status 12 and 13 are outliers, and there are no status 12 items in the selected sample. For status 10 items, 28 of them (2.4%) are regarded as outliers (i.e., 1,166 – 1,138 = 28). For status XX items, eight of them (0.2%) are outliers (i.e., 2,770 – 2,762 = 8). These rates of outliers by calibration status are extremely low, which proves that the quality of the calibration of Spanish math items using Spanish test responses is extremely high.

Table 2.2. Collected and Selected Samples by Test

	Collected Sample		Selected Sample	
Test	#Items	%	#Items	%
Growth: Spanish Math 2–5 AERO 2015	54	1.37	54	1.38
Growth: Spanish Math 2–5 CA 2010	1	0.03	1	0.03
Growth: Spanish Math 2–5 CCSS 2010 V2	553	13.98	552	14.12
Growth: Spanish Math 2–5 FL 2014	23	0.58	23	0.59
Growth: Spanish Math 2–5 General 2019	378	9.56	377	9.65
Growth: Spanish Math 2–5 MI 2010	72	1.82	70	1.79
Growth: Spanish Math 2–5 TX 2012	143	3.62	143	3.66
Growth: Spanish Math 6+ AERO 2015	218	5.51	218	5.58
Growth: Spanish Math 6+ AERO 2015 V2	1	0.03	1	0.03
Growth: Spanish Math 6+ CCSS 2010 V2	303	7.66	298	7.63
Growth: Spanish Math 6+ FL 2014	17	0.43	17	0.44
Growth: Spanish Math 6+ General 2019	62	1.57	62	1.59
Growth: Spanish Math 6+ MI 2010	51	1.29	51	1.31
Growth: Spanish Math 6+ TX 2012	407	10.29	396	10.13
Growth: Spanish Math K–2 CCSS 2010 V2	9	0.23	9	0.23
Growth: Spanish Math K–2 CCSS Intl 2010	27	0.68	27	0.69

Test	Collected Sample		Selected Sample	
	#Items	%	#Items	%
Growth: Spanish Math K–2 FL 2014	807	20.40	792	20.27
Growth: Spanish Math K–2 General 2019	8	0.20	8	0.20
Growth: Spanish Math K–2 MI 2010	37	0.94	37	0.95
Growth: Spanish Math K–2 TX 2012	86	2.17	82	2.1
MAP: Spanish Math 2–5 Common Core 2010	199	5.03	198	5.07
MAP: Spanish Math 2–5 TX 2012	86	2.17	85	2.18
MAP: Spanish Math 6+ Common Core 2010 V	174	4.40	173	4.43
MAP: Spanish Math 6+ TX 2012	239	6.04	234	5.99
Total	3,955	100.00	3,908	100.00

Table 2.3. Collected and Selected Samples by Calibration Status

Calibration Status	Collected Sample		Selected Sample	
	#Items	%	#Items	%
07	8	0.20	8	0.20
10	1,166	29.48	1,138	29.12
12	10	0.25	–	–
13	1	0.03	–	–
XX	2,770	70.04	2,762	70.68
Total	3,955	100.00	3,908	100.00

2.2.2. Person Data

Two types of person data used in the analysis include (1) monolingual data that include student Spanish math test results only and (2) bilingual data that include both Spanish and English math results. The bilingual test results are part of the student Spanish math results in the monolingual data. The collected sample includes valid records pulled from the system, whereas the selected sample only includes selected records according to the following person selection criterion:

> *Absolute difference between old person RIT score (using old Spanish item parameters) and new RIT score (using new Spanish items) is less than or equal to 5 RITs*

Table 2.4 and Table 2.5 present the student demographic information from the collected sample and selected samples for both the monolingual and bilingual data. Table 2.6 and Table 2.7 then present the demographic information of students by grade of the selected sample for monolingual and bilingual data. Of the 147,094 students in the collected sample for the monolingual data, 50 of them (0.03%) were outliers and thus removed from the selected sample. Of the 33,624 students in the collected sample for the bilingual data, 17 of them (0.05%) were outliers and thus removed.

Regardless of monolingual or bilingual data, most students (about 75% across grades) are Spanish students. This is different from the English MAP Growth Math test population in which the most students are white. The major ethnicity difference between the English and Spanish math test populations has important implications in creating Spanish norms and interpreting MAP Growth Math test scores.

Table 2.4. Person Sample Demographics by Collected and Selected Sample—Monolingual Data

Demographic Variable	Collected Sample		Selected Sample	
	N	%	N	%
Total	147,094	100.00	147,044	100.0
Grade				
1	12,456	8.47	12,456	8.47
2	28,906	19.65	28,896	19.65
3	24,197	16.45	24,192	16.45
4	17,491	11.89	17,487	11.89
5	16,202	11.01	16,189	11.01
6	9,700	6.59	9,700	6.6
7	6,803	4.62	6,797	4.62
8	7,004	4.76	7,003	4.76
9	7,042	4.79	7,034	4.78
10	3,620	2.46	3,617	2.46
11	2,110	1.43	2,110	1.43
12	672	0.46	672	0.46
K	10,891	7.4	10,891	7.41
Gender				
Female	72,061	48.99	72,040	48.99
Male	74,814	50.86	74,785	50.86
N/A	219	0.15	219	0.15
Ethnicity				
American Indian or Alaskan	1,921	1.31	1,920	1.31
Asian or Pacific Islander	1,939	1.32	1,939	1.32
Black	1,807	1.23	1,806	1.23
Hispanic	111,112	75.54	111,078	75.54
Native Hawaiian or Other Pacific Islander	133	0.09	133	0.09
White	11,109	7.55	11,103	7.55
Multi-Ethnic	4,215	2.87	4,212	2.86
Not Specified or Other	13,514	9.19	13,509	9.19
N/A	1,344	0.91	1,344	0.91

Table 2.5. Person Sample Demographics by Collected and Selected Sample—Bilingual Data

Demographic Variable	Collected Sample		Selected Sample	
	N	%	N	%
Total	33,624	100.0	33,607	100.0
Grade				
1	3,134	9.30	3,134	9.30
2	6,808	20.30	6,808	20.30
3	5,859	17.40	5,854	17.40
4	4,653	13.80	4,647	13.80
5	4,262	12.70	4,257	12.70
6	1,729	5.10	1,729	5.10
7	1,361	4.10	1,361	4.10
8	1,266	3.80	1,266	3.80
9	849	2.50	848	2.50
10	308	0.90	308	0.90
11	115	0.30	115	0.30
12	57	0.20	57	0.20
K	3,223	9.60	3,223	9.60
Gender				
Female	16,466	49.0	16,455	49.00
Male	17,120	50.9	17,114	50.90
N/A	38	0.10	38	0.10
Ethnicity				
American Indian or Alaskan	437	1.30	437	1.30
Asian or Pacific Islander	354	1.05	354	1.05
Black	842	2.50	841	2.50
Hispanic	26,280	78.16	26,269	78.17
Native Hawaiian or Other Pacific Islander	45	0.13	45	0.13
White	2,854	8.49	2,851	8.48
Multi-Ethnic	604	1.80	602	1.79
Not Specified or Other	2,160	6.42	2,160	6.43
N/A	48	0.14	48	0.14

Table 2.6. Person Sample Demographics by Grade for Selected Sample—Monolingual Data

		%Gender*			%Ethnicity**								
Grade	**N**	**Female**	**Male**	**N/A**	**AI/AN**	**Asian**	**Black**	**Hispanic**	**NH/PI**	**White**	**Multiethnic**	**NS/Other**	**N/A**
K	10,891	49.73	50.23	0.04	1.37	1.02	2.18	77.96	0.06	9.03	2.12	5.54	0.73
1	12,456	50.21	49.74	0.05	1.40	0.92	1.75	76.85	0.11	10.47	2.11	5.82	0.57
2	28,896	50.11	49.84	0.05	1.00	0.61	1.43	77.51	0.12	9.62	1.63	7.86	0.22
3	24,192	49.98	49.96	0.06	1.06	1.30	1.36	74.76	0.09	7.54	2.99	10.43	0.46
4	17,487	49.71	50.08	0.21	1.21	1.33	1.28	76.89	0.06	6.9	2.73	8.86	0.75
5	16,189	49.28	50.51	0.21	1.33	1.44	1.10	76.54	0.04	7.82	2.75	8.26	0.72
6	9,700	47.92	51.73	0.35	2.31	2.59	0.62	75.04	0.12	3.74	3.55	10.38	1.65
7	6,797	47.51	52.21	0.28	1.80	1.53	0.64	70.70	0.08	3.50	4.27	15.49	2.53
8	7,003	45.87	53.81	0.33	2.38	3.27	0.47	69.51	0.10	4.56	2.60	14.55	2.56
9	7,034	43.22	56.57	0.21	0.75	0.40	0.45	74.68	0.11	5.27	5.96	10.88	1.49
10	3,617	45.65	54.02	0.33	0.80	0.19	0.41	74.01	0.17	5.36	6.30	10.45	2.29
11	2,110	48.20	51.47	0.33	0.57	0.28	0.47	64.64	0.05	5.45	8.25	17.39	2.89
12	672	49.26	50.60	0.15	0.60	0.30	0.15	70.98	0.15	17.41	0.30	8.63	1.49

**N/A = Gender information is not available.
**AI/AN = American Indian or Alaskan Native. NH/PI = Native Hawaiian or Other Pacific Islander. NS/Other = Not Specified or Other. N/A= Race and ethnicity information is not available.

Table 2.7. Person Sample Demographics by Grade for Selected Sample—Bilingual Data

		%Gender*			% Ethnicity**								
Grade	**N**	**Female**	**Male**	**N/A**	**AI/AN**	**Asian**	**Black**	**Hispanic**	**NH/PI**	**White**	**Multiethnic**	**NS/Other**	**N/A**
K	3,223	49.36	50.61	0.03	1.27	1.52	3.41	75.36	0.06	10.30	0.90	6.33	0.84
1	3,134	50.86	49.11	0.03	1.05	1.28	2.39	72.27	0.22	11.07	2.52	9.16	0.03
2	6,808	49.66	50.29	0.04	0.98	1.22	2.53	77.25	0.16	8.90	2.50	6.39	0.07
3	5,854	49.66	50.26	0.09	1.47	0.65	3.06	79.02	0.15	9.12	1.40	5.07	0.05
4	4,647	49.13	50.59	0.28	1.40	0.80	2.71	82.03	0.06	7.36	1.29	4.30	0.04
5	4,257	48.74	51.19	0.07	1.57	0.99	1.97	81.04	0.02	9.11	1.39	3.88	0.02
6	1,729	48.64	51.24	0.12	1.04	1.16	1.74	74.26	0.17	4.97	3.35	13.01	0.29
7	1,361	45.78	54.15	0.07	1.10	1.40	1.62	79.43	0.07	3.97	1.84	10.36	0.22
8	1,266	46.68	52.92	0.39	1.50	1.74	1.34	79.30	0.08	4.42	1.58	9.95	0.08
9	848	41.98	57.90	0.12	2.12	0.24	1.89	79.13	0.24	8.61	0.94	6.84	0.00
10	308	44.16	55.19	0.65	1.95	0.65	1.62	79.22	0.97	7.47	3.57	4.55	0.00
11	115	41.74	57.39	0.87	1.74	0.00	4.35	83.48	0.87	5.22	0.00	4.35	0.00
12	57	50.88	49.12	0.00	0.00	0.00	0.00	84.21	1.75	7.02	1.75	5.26	0.00

**N/A = Gender information is not available.
**AI/AN = American Indian or Alaskan Native. NH/PI = Native Hawaiian or Other Pacific Islander. NS/Other = Not Specified or Other. N/A= Race and ethnicity information is not available.

3. Results

3.1. Item Results

3.1.1. Descriptive Statistics

Table 3.1 and Table 3.2 present the descriptive statistics of the old and new RITs and the correlations between them by calibration status (10 and XX only) for the selected sample by calibration status and test, including the mean, standard deviation (SD), n-count, and correlations between the old and new RITs. The distributions of the item RIT mean and SD between the old and new item parameters are very close, with the RIT differences being one decimal point. The correlations between the old and new item parameters are 0.98 across item calibration status.

Table 3.1. Item Descriptive Statistics for Selected Sample by Calibration Status (10 and XX)

Calibration Status	Statistics	Old RIT	New RIT
10	Mean	226.17	226.29
	SD	29.44	29.31
	N	1,138	1,138
	CORR (Old RIT)	1.00	0.98
	CORR (New RIT)	–	1.00
XX	Mean	192.22	192.31
	SD	31.30	31.38
	N	2,762	2,762
	CORR (Old RIT)	1.00	0.98
	CORR (New RIT)	–	1.00

Table 3.2. Item Descriptive Statistics for Selected Sample by Test

Test	Statistics	Old RIT	New RIT
Growth: Spanish Math 2–5 AERO 2015	Mean	209.80	209.98
	SD	19.74	20.58
	N	54.00	54.00
	CORR (Old RIT)	–	0.96
	CORR (New RIT)	0.96	1.00
Growth: Spanish Math 2–5 CA 2010	Mean	186.00	198.00
	SD	–	–
	N	1.00	1.00
	CORR (Old RIT)	–	–
	CORR (New RIT)	–	–
Growth: Spanish Math 2–5 CCSS 2010 V2	Mean	209.50	209.45
	SD	22.53	22.78
	N	552.00	552.00
	CORR (Old RIT)	1.00	0.98
	CORR (New RIT)	–	1.00

Test	Statistics	Old RIT	New RIT
Growth: Spanish Math 2–5 FL 2014	Mean	199.78	201.00
	SD	18.13	19.98
	N	23.00	23.00
	CORR (Old RIT)	1.00	0.95
	CORR (New RIT)	–	1.00
Growth: Spanish Math 2–5 General 2019	Mean	205.91	206.31
	SD	21.61	21.63
	N	377.00	377.00
	CORR (Old RIT)	1.00	0.97
	CORR (New RIT)	–	1.00
Growth: Spanish Math 2–5 MI 2010	Mean	203.36	203.91
	SD	16.90	17.45
	N	70.00	70.00
	CORR (Old RIT)	1.00	0.94
	CORR (New RIT)	–	1.00
Growth: Spanish Math 2–5 TX 2012	Mean	219.04	219.84
	SD	12.52	13.18
	N	143.00	143.00
	CORR (Old RIT)	1.00	0.91
	CORR (New RIT)		1.00
Growth: Spanish Math 6+ AERO 2015	Mean	204.08	204.54
	SD	21.90	22.22
	N	218.00	218.00
	CORR (Old RIT)	1.00	0.97
	CORR (New RIT)	–	1.00
Growth: Spanish Math 6+ CCSS 2010 V2	Mean	232.23	232.00
	SD	29.45	29.38
	N	298.00	298.00
	CORR (Old RIT)	1.00	0.98
	CORR (New RIT)	–	1.00
Growth: Spanish Math 6+ FL 2014	Mean	224.24	222.76
	SD	11.68	16.26
	N	17.00	17.00
	CORR (Old RIT)	1.00	0.87
	CORR (New RIT)	–	1.00
Growth: Spanish Math 6+ General 2019	Mean	244.42	243.94
	SD	11.35	12.43
	N	62.00	62.00
	CORR (Old RIT)	1.00	0.90
	CORR (New RIT)	–	1.00
Growth: Spanish Math 6+ MI 2010	Mean	211.67	211.20
	SD	20.30	19.74
	N	49.00	49.00
	CORR (Old RIT)	1.00	0.96
	CORR (New RIT)	–	1.00

Test	Statistics	Old RIT	New RIT
Growth: Spanish Math 6+ TX 2012	Mean	228.17	228.11
	SD	23.43	23.48
	N	394.00	394.00
	CORR (Old RIT)	1.00	0.97
	CORR (New RIT)	–	1.00
Growth: Spanish Math K–2 CCSS 2010 V2	Mean	207.33	206.33
	SD	11.58	12.78
	N	9.00	9.00
	CORR (Old RIT)	1.00	0.97
	CORR (New RIT)	–	1.00
Growth: Spanish Math K–2 CCSS Intl 2010	Mean	177.48	178.11
	SD	24.81	24.91
	N	27.00	27.00
	CORR (Old RIT)	1.00	0.98
	CORR (New RIT)	–	1.00
Growth: Spanish Math K–2 FL 2014	Mean	155.22	155.61
	SD	22.64	22.44
	N	791.00	791.00
	CORR (Old RIT)	1.00	0.97
	CORR (New RIT)	–	1.00
Growth: Spanish Math K–2 General 2019	Mean	189.13	189.00
	SD	25.95	27.85
	N	8.00	8.00
	CORR (Old RIT)	1.00	0.99
	CORR (New RIT)	–	1.00
Growth: Spanish Math K–2 MI 2010	Mean	174.57	173.46
	SD	16.81	17.34
	N	37.00	37.00
	CORR (Old RIT)	1.00	0.97
	CORR (New RIT)	–	1.00
Growth: Spanish Math K–2 TX 2012	Mean	197.18	197.39
	SD	20.04	20.67
	N	82.00	82.00
	CORR (Old RIT)	1.00	0.96
	CORR (New RIT)	–	1.00
MAP: Spanish Math 2–5 Common Core 2010	Mean	207.64	207.59
	SD	18.16	20.00
	N	197.00	197.00
	CORR (Old RIT)	1.00	0.95
	CORR (New RIT)		1.00
MAP: Spanish Math 2–5 TX 2012	Mean	218.45	220.98
	SD	20.59	20.97
	N	85.00	85.00
	CORR (Old RIT)	1.00	0.96
	CORR (New RIT)	0.96	1.00

Test	Statistics	Old RIT	New RIT
MAP: Spanish Math 6+ Common Core 2010 V	Mean	215.52	213.78
	SD	30.14	31.32
	N	173.00	173.00
	CORR (Old RIT)	1.00	0.98
	CORR (New RIT)	–	1.00
MAP: Spanish Math 6+ TX 2012	Mean	215.01	214.73
	SD	23.98	23.71
	N	232.00	232.00
	CORR (Old RIT)	1.00	0.97
	CORR (New RIT)	–	1.00

3.1.2. Paired Sample T-Test

The paired sample t-test, sometimes called the dependent sample t-test, is a statistical procedure used to determine whether the mean difference between two sets of observations is zero. The null hypothesis (H_0) assumes that the true mean difference (μ_d) is equal to zero. Because the purpose here is to find the mean difference between two calibration values on the same item parameters, the μ_d here is the difference of item RITs between the old and new item parameters. The mathematical representations of the null and alternative hypotheses are defined below:

H_0: $\mu_d = 0$
H_1: $\mu_d \neq 0$ (two-tailed)

Table 3.3 presents the results of paired t-test of the old and new item parameters by item calibration status. All statistical test null hypotheses have been retained, and there are no statistically significant differences between old and new item parameters across item calibration status.

Table 3.3. Paired T-Test by Item Calibration Status for Selected Sample

Pair Sample	Difference	N	Mean	SD	t	DF	P-Value
XX only	Old RIT – New RIT	2,762	-0.09	5.74	-0.84	2761.00	0.40
XX and 10	Old RIT – New RIT	3,900	-0.10	5.71	-1.09	3899.00	0.28

3.2. Person Results

The impact of item parameters on student Spanish math scores can be investigated by comparing student scores from different Spanish math item parameters. In this study, each student has two different Spanish math RIT scores:

1. Old student score (Spanish RIT) based on the old Spanish math transadapted item parameters (old RIT) that are equal to their English counterparts
2. New student score (Spanish NRIT) based on the newly calibrated Spanish item parameters (new RIT)

As shown previously in Table 2.3, about 30% of newly calibrated items are labeled as calibration status 10, which can still be used operationally in the Spanish math test. Some English transadapted items with calibration status XX in the English math tests may have calibration status 10 as newly calibrated Spanish items. To investigate the impact of treating these status 10 items on student scores, the first method is to treat these status 10 items from new calibration as operational items (XX) in scoring. The second method is to treat these status 10 items as field test items and keep their item parameters as the old Spanish item parameters in scoring.

Table 3.4 presents the descriptive statistics and correlations of student scores by grade of the selected sample for both the monolingual and bilingual data based on items with calibration status 10, whereas Table 3.5 presents the descriptive statistics for the monolingual data only based on items without calibration status 10. Similarly, Table 3.6 presents the descriptive statistics and correlations of student scores by ethnicity of the selected sample for both the monolingual and bilingual data based on items with calibration status 10, whereas Table 3.7 presents the descriptive statistics for the monolingual data only based on items without calibration status 10.

For both datasets by grade and ethnicity, the distributions of the person RIT mean and SD between old and new item parameters (i.e., between Spanish RIT and Spanish NRIT) are very close, with the RIT differences being one decimal point. The correlations between old and new item parameters are 1.00 across grades. The difference between the monolingual results based on items with and without item calibration status 10 is very small for both grade and ethnicity, which indicates that the impact of treating the newly calibrated status 10 items as status XX items is very small and can be neglectable.

Table 3.4. Person Descriptive Statistics for Selected Sample by Grade—Items with Calibration Status 10

		Monolingual Data		Bilingual Data	
Grade	**Statistics**	**Spanish RIT**	**Spanish NRIT**	**Spanish RIT**	**Spanish NRIT**
K	Mean	137.29	137.56	138.02	138.37
	SD	14.14	13.85	14.21	13.88
	N	10,891	10,891	3,223	3,223
	CORR (Spanish RIT)	1.00	1.00	1.00	1.00
	CORR (Spanish NRIT)	–	1.00	–	1.00
1	Mean	157.76	157.68	159.03	159.01
	SD	17.33	17.19	18.95	18.74
	N	12,456	12,456	3,134	3,134
	CORR (Spanish RIT)	1.00	1.00	1.00	1.00
	CORR (Spanish NRIT)	–	1.00	–	1.00
2	Mean	180.32	180.24	179.41	179.39
	SD	14.76	14.51	15.37	15.09
	N	28,896	28,896	6,808	6,808
	CORR (Spanish RIT)	1.00	1.00	1.00	1.00
	CORR (Spanish NRIT)	–	1.00	–	1.00

Grade	Statistics	Monolingual Data		Bilingual Data	
		Spanish RIT	Spanish NRIT	Spanish RIT	Spanish NRIT
3	Mean	192.79	192.73	191.89	191.87
	SD	14.77	14.60	15.27	15.08
	N	24,192	24,192	5,854	5,854
	CORR (Spanish RIT)	1.00	1.00	1.00	1.00
	CORR (Spanish NRIT)	–	1.00	–	1.00
4	Mean	202.88	202.78	202.21	202.16
	SD	16.74	16.53	17.44	17.24
	N	17,487	17,487	4,647	4,647
	CORR (Spanish RIT)	1.00	1.00	1.00	1.00
	CORR (Spanish NRIT)	–	1.00	–	1.00
5	Mean	210.64	210.47	209.99	209.85
	SD	18.08	17.84	17.92	17.73
	N	16,189	16,189	4,257	4,257
	CORR (Spanish RIT)	1.00	1.00	1.00	1.00
	CORR (Spanish NRIT)	–	1.00	–	1.00
6	Mean	209.82	209.73	206.61	206.51
	SD	16.57	16.32	16.36	16.23
	N	9,700	9,700	1,729	1,729
	CORR (Spanish RIT)	1.00	1.00	1.00	1.00
	CORR (Spanish NRIT)	–	1.00	–	1.00
7	Mean	210.86	210.82	207.32	207.24
	SD	17.38	17.14	17.19	17.08
	N	6,797	6,797	1,361	1,361
	CORR (Spanish RIT)	1.00	1.00	1.00	1.00
	CORR (Spanish NRIT)	–	1.00	–	1.00
8	Mean	214.56	214.41	212.65	212.44
	SD	19.00	18.71	18.80	18.66
	N	7,003	7,003	1,266	1,266
	CORR (Spanish RIT)	1.00	1.00	1.00	1.00
	CORR (Spanish NRIT)	–	1.00	–	1.00
9	Mean	215.13	214.94	208.80	208.66
	SD	20.73	20.50	20.39	20.16
	N	7,034	7,034	848	848
	CORR (Spanish RIT)	1.00	1.00	1.00	1.00
	CORR (Spanish NRIT)	–	1.00	–	1.00
10	Mean	220.07	219.86	209.55	209.49
	SD	21.01	20.74	17.16	17.03
	N	3,617	3,617	308	308
	CORR (Spanish RIT)	1.00	1.00	1.00	1.00
	CORR (Spanish NRIT)	–	1.00	–	1.00

		Monolingual Data		Bilingual Data	
Grade	Statistics	Spanish RIT	Spanish NRIT	Spanish RIT	Spanish NRIT
11	Mean	226.71	226.39	211.97	212.01
	SD	22.34	22.03	18.41	18.29
	N	2,110	2,110	115	115
	CORR (Spanish RIT)	1.00	1.00	1.00	1.00
	CORR (Spanish NRIT)	–	1.00	–	1.00
12	Mean	222.57	222.38	212.51	212.54
	SD	21.23	20.97	17.16	17.16
	N	672	672	57	57
	CORR (Spanish RIT)	1.00	1.00	1.00	1.00
	CORR (Spanish NRIT)	–	1.00	–	1.00

Table 3.5. Person Descriptive Statistics for Selected Sample by Grade—Items without Calibration Status 10

		Monolingual Data	
Grade	Statistics	Spanish RIT	Spanish NRIT
K	Mean	137.28	137.40
	SD	14.13	13.88
	N	10,888	10,888
	CORR (Spanish RIT)	1.00	1.00
	CORR (Spanish NRIT)	–	1.00
1	Mean	157.75	157.64
	SD	17.32	17.21
	N	12,450	12,450
	CORR (Spanish RIT)	1.00	1.00
	CORR (Spanish NRIT)	–	1.00
2	Mean	180.33	180.22
	SD	14.77	14.45
	N	28,917	28,917
	CORR (Spanish RIT)	1.00	1.00
	CORR (Spanish NRIT)	–	1.00
3	Mean	192.79	192.72
	SD	14.76	14.59
	N	24,203	24,203
	CORR (Spanish RIT)	1.00	1.00
	CORR (Spanish NRIT)	–	1.00
4	Mean	202.89	202.76
	SD	16.72	16.51
	N	17,511	17,511
	CORR (Spanish RIT)	1.00	1.00
	CORR (Spanish NRIT)	–	1.00

Grade	Statistics	Monolingual Data	
		Spanish RIT	Spanish NRIT
5	Mean	210.65	210.45
	SD	18.10	17.85
	N	16,173	16,173
	CORR (Spanish RIT)	1.00	1.00
	CORR (Spanish NRIT)	–	1.00
6	Mean	209.84	209.75
	SD	16.58	16.32
	N	9,710	9,710
	CORR (Spanish RIT)	1.00	1.00
	CORR (Spanish NRIT)	–	1.00
7	Mean	210.85	210.81
	SD	17.39	17.14
	N	6,792	6,792
	CORR (Spanish RIT)	1.00	1.00
	CORR (Spanish NRIT)	–	1.00
8	Mean	214.56	214.42
	SD	19.00	18.74
	N	7,007	7,007
	CORR (Spanish RIT)	1.00	1.00
	CORR (Spanish NRIT)	–	1.00
9	Mean	215.12	214.96
	SD	20.76	20.56
	N	7,050	7,050
	CORR (Spanish RIT)	1.00	1.00
	CORR (Spanish NRIT)	–	1.00
10	Mean	220.07	219.90
	SD	21.04	20.84
	N	3,622	3,622
	CORR (Spanish RIT)	1.00	1.00
	CORR (Spanish NRIT)	–	1.00
11	Mean	226.68	226.45
	SD	22.37	22.20
	N	2,102	2,102
	CORR (Spanish RIT)	1.00	1.00
	CORR (Spanish NRIT)	–	1.00
12	Mean	222.47	222.32
	SD	21.18	21.04
	N	669	669
	CORR (Spanish RIT)	1.00	1.00
	CORR (Spanish NRIT)	–	1.00

Table 3.6. Person Descriptive Statistics for Selected Sample by Ethnicity—Items with Calibration Status 10

		Monolingual Data		Bilingual Data	
Ethnicity	**Statistics**	**Spanish RIT**	**Spanish NRIT**	**Spanish RIT**	**Spanish NRIT**
American Indian or Alaskan	Mean	189.99	189.97	188.20	188.18
	SD	26.51	26.25	27.12	26.86
	N	1,920	1,920	437	437
	CORR (Spanish RIT)	1.00	1.00	1.00	1.00
	CORR (Spanish NRIT)	–	1.00	1.00	1.00
Asian or Pacific Islander	Mean	186.60	187.08	180.03	180.28
	SD	24.59	24.49	31.26	30.97
	N	1,939	1,939	354	354
	CORR (Spanish RIT)	1.00	1.00	1.00	1.00
	CORR (Spanish NRIT)	–	1.00	1.00	1.00
Black	Mean	175.73	175.73	175.50	175.50
	SD	29.98	29.65	29.45	29.12
	N	1,806	1,806	841	841
	CORR (Spanish RIT)	1.00	1.00	1.00	1.00
	CORR (Spanish NRIT)	–	1.00	1.00	1.00
Hispanic	Mean	190.74	190.67	188.53	188.51
	SD	27.80	27.61	27.43	27.22
	N	111,078	111,078	26,269	26,269
	CORR (Spanish RIT)	1.00	1.00	1.00	1.00
	CORR (Spanish NRIT)	–	1.00	1.00	1.00
Native Hawaiian or Other Pacific Islander	Mean	188.09	188.36	181.62	182.05
	SD	25.34	25.26	26.51	26.35
	N	133	133	45	45
	CORR (Spanish RIT)	1.00	1.00	1.00	1.00
	CORR (Spanish NRIT)	–	1.00	1.00	1.00
White	Mean	190.09	190.03	185.85	185.87
	SD	29.42	29.20	29.28	29.10
	N	11,103	11,103	2,851	2,851
	CORR (Spanish RIT)	1.00	1.00	1.00	1.00
	CORR (Spanish NRIT)	–	1.00	1.00	1.00
Multi-Ethnic	Mean	204.03	203.99	187.85	187.98
	SD	30.23	30.01	28.44	28.20
	N	4,212	4,212	602	602
	CORR (Spanish RIT)	1.00	1.00	1.00	1.00
	CORR (Spanish NRIT)	–	1.00	1.00	1.00
Not Specified or Other	Mean	199.70	199.43	186.87	186.74
	SD	28.74	28.42	27.93	27.61
	N	13,509	13,509	2,160	2,160
	CORR (Spanish RIT)	1.00	1.00	1.00	1.00
	CORR (Spanish NRIT)	–	1.00	1.00	1.00

		Monolingual Data		Bilingual Data	
Ethnicity	Statistics	Spanish RIT	Spanish NRIT	Spanish RIT	Spanish NRIT
N/A	Mean	206.35	206.65	154.17	154.56
	SD	33.22	33.06	34.18	34.06
	N	1,344	1,344	48	48
	CORR (Spanish RIT)	1.00	1.00	1.00	1.00
	CORR (Spanish NRIT)	–	1.00	1.00	1.00

Table 3.7. Person Descriptive Statistics for Selected Sample by Ethnicity—Items without Item Calibration Status 10

		Monolingual Data	
Ethnicity	Statistics	Spanish RIT	Spanish NRIT
American Indian or Alaskan	Mean	189.99	189.94
	SD	26.52	26.27
	N	1,920	1,920
	CORR (Spanish RIT)	1.00	1.00
	CORR (Spanish NRIT)	–	1.00
Asian or Pacific Islander	Mean	186.60	187.05
	SD	24.58	24.50
	N	1,939	1,939
	CORR (Spanish RIT)	1.00	1.00
	CORR (Spanish NRIT)	–	1.00
Black	Mean	175.74	175.67
	SD	29.98	29.68
	N	1,806	1,806
	CORR (Spanish RIT)	1.00	1.00
	CORR (Spanish NRIT)	–	1.00
Hispanic	Mean	190.74	190.64
	SD	27.80	27.64
	N	111,081	111,081
	CORR (Spanish RIT)	1.00	1.00
	CORR (Spanish NRIT)	–	1.00
Native Hawaiian or Other Pacific Islander	Mean	188.09	188.34
	SD	25.34	25.25
	N	133	133
	CORR (Spanish RIT)	1.00	1.00
	CORR (Spanish NRIT)	–	1.00
White	Mean	190.10	189.98
	SD	29.42	29.21
	N	11,106	11,106
	CORR (Spanish RIT)	1.00	1.00
	CORR (Spanish NRIT)	–	1.00

Ethnicity	Statistics	Monolingual Data	
		Spanish RIT	Spanish NRIT
Multi-Ethnic	Mean	204.03	203.94
	SD	30.24	30.03
	N	4,212	4,212
	CORR (Spanish RIT)	1.00	1.00
	CORR (Spanish NRIT)	–	1.00
Not Specified or Other	Mean	199.70	199.44
	SD	28.74	28.50
	N	13,509	13,509
	CORR (Spanish RIT)	1.00	1.00
	CORR (Spanish NRIT)	–	1.00
N/A	Mean	206.35	206.56
	SD	33.21	33.06
	N	1,344	1,344
	CORR (Spanish RIT)	1.00	1.00
	CORR (Spanish NRIT)	–	1.00

4. Conclusion

4.1. Summary of Results

Because all Spanish math tests are adaptive, the calibration was conducted for a pool of items, not items from a traditional linear test. For adaptive item calibration, due to student ability difference, the distribution of the number of student responses collected for a pool items is a uniform distribution as in a linear test. This is why about 1/3 calibrated Spanish math items have calibration status 10, and even all these status 10 items have status XX when they were calibrated in English math tests.

To answer the first research question regarding the calibration impact on new item parameters when comparing to the old item parameters, evidences on the difference between the old and new RIT item parameters show that this difference is very small. Treating item status 10 items as status XX items in the Spanish math test does not affect the differences between the old and new item parameters. The differences of means and SDs of item parameters by either item calibration status or test between the old and new calibrations are in one RIT decimal point, and correlations between the old and new RITs is 0.98 across item calibration status and test. The dependent t-test results show that there are no statistically significant differences between the old and new RITs across item calibration status and test.

The second and third research questions concern about the impact of item parameter calibration on Spanish student math scores for monolingual student data. Based on the evidence from this study, regardless of including status 10 items in student scoring, the student person parameters of Spanish RIT and Spanish NRIT are almost the same; the differences of means and SDs between Spanish RIT and Spanish NRIT for items both with and without item status 10 items in person scoring are in one RIT decimal point; and correlations between the old and new RITs is 1.00 by grade and ethnicity. The implication of no impact of item calibration status 10 on student scores is that all item status 10 items from the Spanish math calibration process can be treated as status XX items. If they are treated as status 10 items, NWEA needs to accumulate more student responses and calibrate them later.

The fourth research question is similar to the second and third research questions, except that the student data is bilingual data and all scoring methods include all status 10 items. The evidence shows that the person parameters of Spanish RIT and Spanish NRIT are almost the same; the differences of means and SDs between Spanish RIT and Spanish NRIT are in one RIT decimal point; and correlations between the old and new RITs is 1.00 by grade and ethnicity.

Another interesting finding from this study is the major ethnicity difference between the English and Spanish math test populations. This has important implications in creating Spanish norms and interpreting MAP Growth Math test scores. For example, the English math test contains about 15% Hispanic students, whereas the Spanish math test contains about 78% Hispanic students. The Spanish norms could be developed using the Spanish scale, which would increase the relevance of the norms to allow Hispanic students to be compared to other Hispanic students.

4.2. Recommendations

The major finding of this study is that instead of borrowing English math item parameters for the Spanish math tests, calibrating Spanish math item parameters using students' empirical responses is a much better approach to constructing the Spanish MAP Growth Math tests according to psychometric scaling and validity theory. Our recommendations for operational practice of Spanish math test item calibration are as follows:

1. Replace all current English math item parameters used in the Spanish math tests with calibrated Spanish math item parameters.
2. Stop using English math item parameters in the Spanish math tests in the future.
3. Implement NWEA routine calibration procedures for all Spanish math items in the future.
4. For all item calibration status 12 and 13 items in Spanish item calibration, ask NWEA content experts to review them and exclude them now from Spanish math item banks before final decision made by the content experts.
5. For all item calibration status 10 items in Spanish item calibration, keep using them as operational items and review them when calibration n-counts are large enough.
6. Conduct item parameter drift study once a year after replacement has been made.