Article

Journal of Psychoeducational Assessment
1–16

DOI: 10.1177/0734282918809793
journals.sagepub.com/home/jpa
SAGE

Psychometric Properties of the WRAT Math Computation Subtest in Mexican Adolescents

Roberto A. Abreu-Mendoza[1], Yaira Chamorro[1], and Esmeralda Matute[1]

Abstract

The goal of this study was to provide normative scores and examine the psychometric properties of the Math Computation subtest of the Wide Range Achievement Test–IV (WRAT-IV) for Mexican adolescents after the completion of junior high school. We group-administered this subtest to 1,318 first-year Mexican high school students. We then obtained its overall internal reliability and examined its underlying factor structure. Finally, we determined its concurrent and criterion validity by evaluating a subsample of 106 students that included adolescents with mathematical difficulty, mathematical talent, and typical performance. Results showed that the subtest has a good internal reliability and appropriate psychometric characteristics, suggesting its appropriateness for the detection of adolescents with particular difficulty or ability in mathematics. The exploratory factor analysis identified three factors: arithmetic, fractions and basic algebra, and rational numbers. There were also sex differences in the number of correct responses, but the effect size was small.

Keywords

validity, factor analysis, mathematics, high school

In Mexico, a large percentage of adolescents do not achieve appropriate mathematical literacy levels by the end of junior high school, which could affect their adult socioeconomic status (Ritchie & Bates, 2013). According to the 2012 Program for International Student Assessment (PISA), 46.8% of 15-year-old Mexican high school students cannot "employ basic algorithms, formulae, procedures, or conventions" (Organisation for Economic Co-operation and Development [OECD], 2014, p. 61), that is, they reached Level 1, but not Level 2, of that assessment (INEE, 2015). Based on these data, it would be difficult to distinguish adolescents with mathematical difficulty from those with low performance because of their educational background. Standardized assessments help to determine whether an individual's score is like or unlike those of a group of individuals with similar characteristics (McCauley & Swisher, 1984); however, there are few such tools for the Mexican population, especially for adolescents.

[1]Universidad de Guadalajara, Mexico

Corresponding Author:
Esmeralda Matute, Instituto de Neurociencias, CUCBA, Universidad de Guadalajara, Francisco de Quevedo 180, CP 44130 Guadalajara, Mexico.
Email: ematute@redudg.udg.mx

One available assessment is the *Evaluación Neuropsicológica Infantil* (ENI [Neuropsychological Assessment for Children]; Matute, Rosselli, Ardila, & Ostrosky, 2007). This individually administered neuropsychological battery was standardized with 800 monolingual Spanish-speaking children, aged 6 to 16 years, from Mexico and Colombia, and comprises several subtests to evaluate different cognitive domains and academic abilities. Its numerical subtests have shown the ability to distinguish between 11- and 12-year-old children with dyscalculia and those with typical performance (Rosselli, Matute, Pinto, & Ardila, 2006), and it has demonstrated appropriate concurrent validity (Rosselli, Ardila, Matute, & Inozemtseva, 2009). Most of these subtests, however, evaluate only basic numerical abilities (e.g., counting and reading numbers). Only the Written Math subtest, which contains arithmetic problems with fractions and decimals, and algebraic problems that involve solving linear equations, captures individual differences in more advanced abilities.

To detect individuals who may have particular difficulty or proficiency in mathematics, it is necessary to have assessments that can be group administered, so they can be used as screening tests to evaluate large samples. The Math Computation subtest from the Wide Range Achievement Test–IV (WRAT-IV; Wilkinson & Robertson, 2006) has been widely used for this purpose. For older children, adolescents, and adults, this subtest consists of 40 arithmetic and algebra problems that have to be solved in 15 min. This subtest has been used not only to evaluate children but also to evaluate individual differences in mathematical abilities of adolescents and adults (Buelow & Frakey, 2013; Starr, DeWind, & Brannon, 2017). However, the current educational context and cultural factors make it difficult to use its published norms for the Mexican population (American Educational Research Association, National Council on Measurement in Education, & American Psychological Association, 2014).

The current study has two primary objectives. The first is to provide normative data and show the internal reliability and concurrent validity of the Math Computation subtest (Wilkinson & Robertson, 2006) for Mexican adolescents in the first year of high school. The second objective is to describe the underlying factor structure of the WRAT Math Computation subtest; to the best of our knowledge, there is no study describing this structure using exploratory factor analyses (EFAs). These possible underlying factors may allow a more fine-grained measurement of adolescents' mathematical abilities than the global number of correct responses. In addition, we had three secondary objectives: (a) to show the criterion validity (Ciccheti, 1994) of this subtest for the categories of mathematical difficulty and mathematical talent; (b) to evaluate possible differences between the performance of boys and girls on this subtest, as adolescence has been a critical developmental stage in the evaluation of sex differences in mathematical abilities (Hyde, Fennema, & Lamon, 1990; Stoet & Geary, 2015); and (c) to investigate the psychometric properties of this subtest using a Rasch model (Rasch, 1960).

Method

Participants

The final sample consisted of 1,318 students (M age = 15.89 years, SD = 0.51 years, 806 female) from two public high schools in Guadalajara, Mexico. Students were in their first year, either in the first (n = 677) or second semester (n = 641), ranging in age from 14.65 to 16.99 years. After administering the WRAT Math Computation subtest, we sent 466 invitation letters seeking to recruit the largest possible number of participants at both extremes of the mathematical ability distribution; however, only 106 families responded to our letter. The resulting subsample of 106 students (M age = 15.90 years, SD = 0.44 years, 69 female) was evaluated an average of 10.09 weeks (SD = 1.35 weeks) after the subtest administration to determine its concurrent validity and other psychometric properties. Table 1 shows the sociodemographic characteristics of the whole

Table 1. Sociodemographic Characteristics of the Whole and Validity Samples.

	Whole sample					
	First semester			Second semester		
	Boys	Girls	All	Boys	Girls	All
	M (*SD*)	*M* (*SD*)	*M* (*SD*)	*M* (*SD*)	*M* (*SD*)	*M* (*SD*)
n	264	413	677	248	393	641
Age	15.60 (0.46)	15.65 (0.44)	15.63 (0.45)	16.13 (0.41)	16.18 (0.42)	16.16 (0.42)
	Validity sample					
	First semester			Second semester		
	Boys	Girls	All	Boys	Girls	All
	M (*SD*)	*M* (*SD*)	*M* (*SD*)	*M* (*SD*)	*M* (*SD*)	*M* (*SD*)
n	28	49	77	9	20	29
Age	15.72 (0.38)	15.80 (0.42)	15.77 (0.41)	16.26 (0.18)	16.27 (0.33)	16.27 (0.29)

and of the validity samples. The whole sample had a uniform female/male ratio and age; however, there was a larger number of first-semester students in the validity sample (73%) than in the whole (51%).

Maternal education was used as a proxy for the socioeconomic status of our sample (Hoff, 2006). Based on data for 442 of the students' mothers, the data were as follows: 1.6% of mothers did not complete elementary school, 12.4% completed elementary school, 25.2% had at least some middle school, 27.8% had at least some high school, and the remaining 33% had at least 1 year of college. Most of the mothers (60.8%) had more than the mean educational level for women in Mexico, which is approximately 9 years of schooling (Instituto Nacional de Estadística y Geografía, 2015). The distribution of maternal education in the validity sample was similar to that of the whole sample: 3.5%, 8.8%, 21.1%, 36.8%, and 29.8%, respectively.

Adolescents older than 17 years old were excluded ($n = 65$) because they were likely to have failed one school year; the usual age of entrance to high school in Mexico is 15. An additional 63 students were excluded because they did not provide complete identification information ($n = 63$): date of birth, age, and sex.

Materials

WRAT Math Computation subtest. This timed subtest and the Word Reading, Sentences Comprehension, and Spelling subtests make up the WRAT-IV (Wilkinson & Robertson, 2006). The test was standardized in the United States with a sample of 3,021 participants, aged 5 to 94 years. The WRAT-IV has two forms; for this study, we administered only the Blue form. As with other standardized tests, raw scores for each subtest and age group can be transformed into standard scores, in this case, ranging from 55 to 145, with a mean standard score of 100 and a standard deviation of 10. Raw scores can also be converted to Rasch ability scaled scores, which the authors recommend for the study of changes over time. For the Math Computation subtest, only a total score is given; that is, no specific subdomains are identified.

The Blue form of the Math Computation subtest consists of 40 items in which participants have to solve written mathematical problems of increasing difficulty, with a 15-min time limit. It

begins with a simple addition problem and ends with the reduction of a rational expression to its lowest terms. The internal consistency for the U.S. sample ranges from .90 to .91 for the age groups 13 to 14 and 15 to 16 years, and its concurrent validity *r* values with similar math subtests range from .50 to .96.

Adaptation. Mexican native Spanish speakers translated the instructions and three verbal items of the WRAT Math Computation subtest (Blue form). Mathematical notation was changed to that used in Mexico: the long division symbol) ¯¯¯¯¯ was changed to ⟌, and fractions were changed from a vertical format to a horizontal one.

Scoring. The WRAT response sheet was used to score the subtest; however, participants were not penalized if they wrote the correct response in a different format (e.g., decimal instead of fraction) or if they did not reduce it to its simplest form. The final score was the total number of correct responses.

Concurrent Measures

To measure the correlation of this adaptation of the WRAT Math Computation subtest with individually administered, Mexican standardized instruments measuring mathematical skills, we employed the following subtests:

ENI Written Math subtest. This subtest of the standardized ENI (Matute et al., 2007) was used to evaluate adolescents' ability to solve written mathematical problems. Participants were evaluated individually and had a maximum of 10 min to answer 14 problems of increasing difficulty. The simplest items involved single-digit arithmetic, whereas the most difficult items involved the addition of like fractions and solving a simple linear equation (e.g., $4x + 2 = 10$). One point was given for each correct answer. For children and adolescents in the age range of 6 years to 16 years 11 months, two scaled scores can be obtained: one for the number of correct answers (accuracy score) and another based on the number of correct answers and the time taken (speed score). Within our sample, the reliability for this subtest was appropriate (Cronbach's $\alpha = .76$).

Wechsler Intelligence Scale for Children (WISC) Arithmetic subtest. This subtest of the Mexican version of the WISC-IV (Wechsler, 2007) was used to assess the ability to solve mathematical problems presented orally. The reported internal consistency for this subtest is appropriate (Cronbach's $\alpha = .91$; Wechsler, 2007).

Descriptive Measures

ENI Reading a Text Aloud subtest. Reading accuracy, comprehension, and speed were assessed by the Reading a Text Aloud subtest of the ENI (Matute et al., 2007). In this subtest, participants are asked to read aloud a 101-word text, *Tontolobo y el Carnero* ("The Silly Wolf and the Ram"); they are then asked four text-comprehension questions. For children in the age range of 7 years to 16 years 11 months, scaled scores can be obtained for reading accuracy, comprehension, and speed. The reliability for the reading comprehension items within our sample was low (Cronbach's $\alpha = .59$).

Estimated IQ. IQ was estimated using a short form of the WISC-IV, which included the Vocabulary and Matrix Reasoning subtests. According to Sattler (2010), this form is one of the 10 best subtest combinations for IQ estimation and has high reliability (.93) and validity (.87) scores.

Procedure

The Math Computation subtest was administered in 2015 and 2016 as a school activity within the first 3 weeks of the semester. The examiner read the instructions to the whole classroom and gave students 15 min to answer the subtest. To obtain the concurrent validity, we convenience sampled 466 students: those whose scores were 1.5 *SD* or more below the mean, those whose scores were 1.5 *SD* or more above the mean, and those whose scores were between the 50th and 60th percentiles. Invitation letters were sent to parents requesting consent to evaluate their children with other subtests from norm-referenced cognitive and academic achievement tests. Only 106 parents (22.75%) agreed; all these provided written informed consent. Evaluation took place in a quiet room at the school and lasted approximately 45 min. Subtest administration was in the following fixed order: Vocabulary, Matrix Reasoning, Arithmetic, Reading a Text Aloud, and Written Math. This study was approved by the university's ethics committee, in accordance with the principles of the Helsinki Declaration.

Results

Normative Scores

The mean score on the Math Computation subtest in our sample was 25.58 ($SD = 5.66$); the skewness and kurtosis were 0.01 and -0.64, respectively. A frequency histogram was constructed using Sturges' rule (Sturges, 1926) to determine the number of bins. Figure 1 shows the histogram and the respective values for important cutoff points reported in the literature.

Exploring the Underlying Factor Structure

Following the guidelines outlined by Preacher and MacCallum (2003), an EFA was used to analyze the underlying factors in the WRAT Math Computation subtest using the "psych" package (Revelle, 2017) in R, the programming language for statistical computing. As a first step, we removed 82 participants who were detected as multivariate outliers using Mahalanobis distance, $\chi^2(40) = 73.40$. Bartlett's test indicated correlation adequacy, $\chi^2(561) = 6{,}769.76$, $p < .001$, and the Kaiser–Meyer–Olkin test indicated measure of sampling adequacy = 0.92.

After eliminating these 82 participants, six items had zero variance and were removed: Items 1 to 4 and Item 10 were answered correctly by all participants, whereas no participant answered Item 40 correctly. Due to the binary nature of the data, robust weighted least squares, which uses polychoric correlations, was used as the estimating procedure with direct oblimin rotation because of expected factor correlation (Schmitt, 2011). A parallel analysis (Horn, 1965) suggested a six-factor model,[1] whereas the scree plot (Catell, 1966) showed the elbow of the curve between two and three eigenvalues, suggesting that a useful model for these data may have either two or three factors (see Figure 2).

A consistent result was obtained with the Jolliffe (1986) criterion, a less conservative alternative to Kaiser's that includes factors with eigenvalues $\geqslant 0.7$. Use of this criterion indicated two factors; however, as can be seen from the scree plot, the eigenvalues for two and three factors were similar (0.77 and 0.61, respectively), and then dropped (four factors = 0.25). We, thus, decided to try both models. The two were similar: The same items loaded for Factors 1 and 2; the main difference was that items that did not load for the two-factor model did load into the third factor of the three-factor model. As this third factor evaluated arithmetic problems that did not involve fractions, we settled on the three-factor model, because at this academic stage, it is theoretically important to distinguish between these two types of arithmetic problems (Siegler & Lortie-Forgues, 2017).

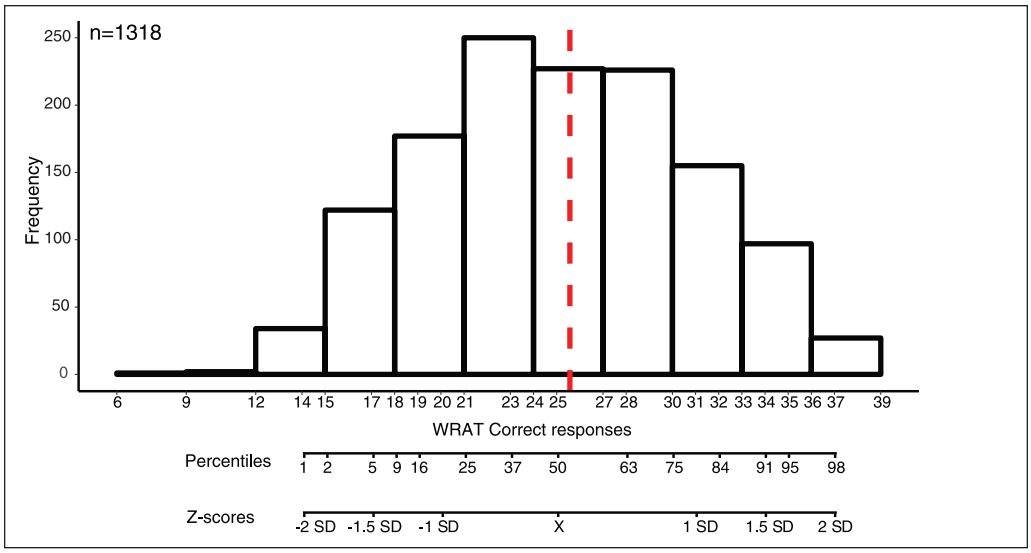

Figure 1. Histogram of the distribution of the correct responses of the total sample in the WRAT Math Computation subtest.
Note. The red (gray) dotted line represents the nonrounded mean. WRAT = Wide Range Achievement Test.

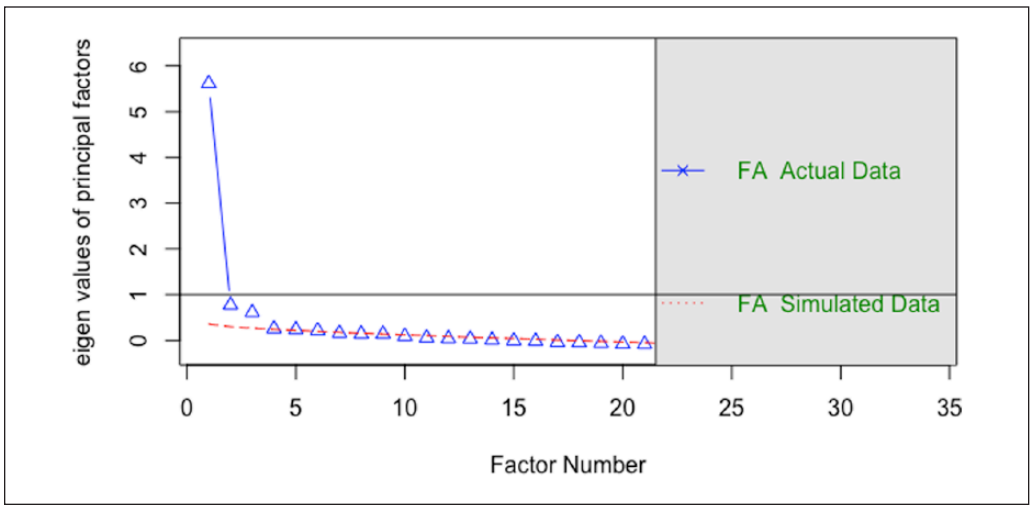

Figure 2. Scree plot of the parallel analysis for the WRAT math computation subtest.
Note. WRAT = Wide Range Achievement Test; FA = factor analysis.

Testing the remaining 34 items showed that 12 items (Items 5-15, 23, and 34) did not load to any factor using the criterion that loadings must be greater than 0.300: Their factor loading ranged from −0.09 to 0.27. These 12 items were eliminated from further analyses. Finally, another three-factor model (see Figure 3) was tested with the remaining 22 items. This model achieved structural simplicity, with each item loading on one and only one factor, and it had excellent fit. The root mean square error of approximation (RMSEA) indicated excellent fit at 0.027, 90% confidence interval (CI) = [0.02, 0.03], and the root mean square of residuals also had excellent fit at 0.02, comparative fit index = 0.97, and Tucker–Lewis index (TLI) = 0.96.

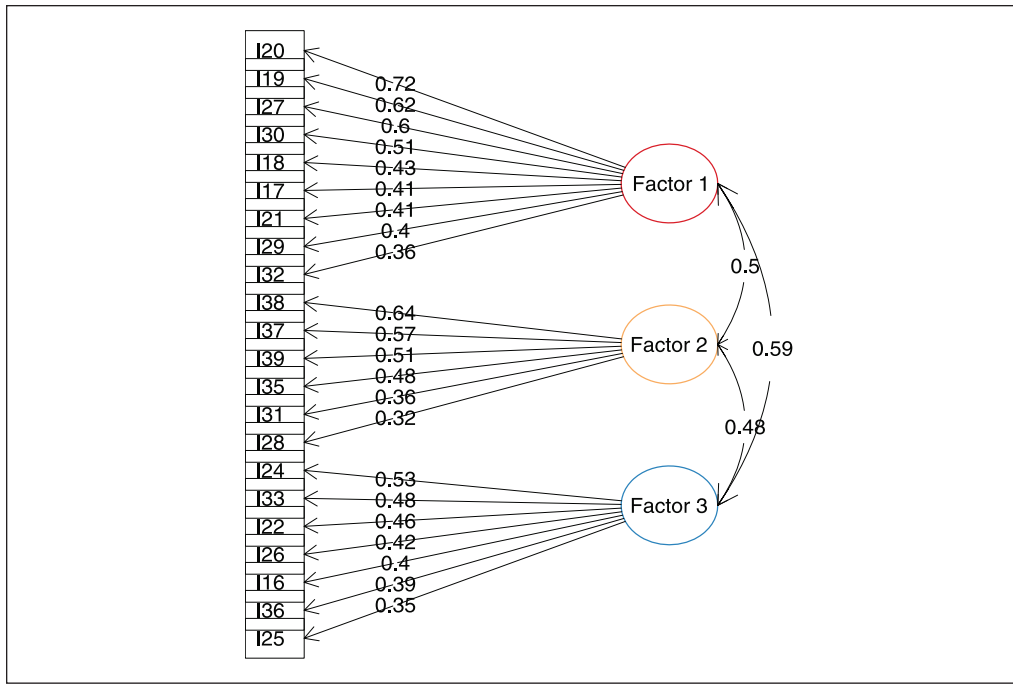

Figure 3. Graphical representation of the final model estimates.

Factor 1, labeled fractions and basic algebra, included nine items in which participants had to solve arithmetic problems of like fractions and simple equations (e.g., $31 + x = 50$). The mean proportion of correct responses was 0.54 ($SD = 0.30$). Factor 2, labeled rational numbers, included six items that involved transforming decimals to percentages or to fractions, and solving arithmetic problems with proper fractions. The mean proportion of correct responses was 0.18 ($SD = 0.24$). Factor 3, labeled arithmetic, included seven items that involved addition, subtraction, multiplication, and division of multidigit numbers with and without decimals. The mean proportion of correct responses was 0.62 ($SD = 0.25$).

Internal Reliability

The overall internal consistency of the Math Computation subtest was good: Cronbach's $\alpha = .85$ (Ciccheti, 1994). The internal consistency of the three factors was as follows: the fraction and basic algebra factor was good ($\alpha = .81$), the rational numbers factor was fair ($\alpha = .70$), and the arithmetic factor was low ($\alpha = .62$).

Academic Semester and Sex Effects

Table 2 shows the mean correct responses in the Math Computation subtest by semester and sex. A two-way ANOVA with semester and sex as between-subjects factors indicated an effect of sex on the number of correct responses in the WRAT Math Computation subtest, $F(1, 1{,}314) = 37.52$, $p < .001$. On average, boys ($M = 26.76$, $SD = 5.78$) outperformed girls ($M = 24.83$, $SD = 5.46$); however, the effect size of these sex differences, as indicated by Cohen's d, was small ($d = 0.34$, 95% CI = [0.23, 0.45]). There was no effect of semester, $F(1, 1{,}314) = 0.85$, $p = .356$, or interaction between these two factors, $F(1, 1314) = 0.56$, $p = .454$.

Table 2. Mean (Standard Deviations) Correct Responses by Semester and Sex of the Whole Sample and of the Validity Sample.

	Whole sample					
	First semester			Second semester		
	Boys	Girls	All	Boys	Girls	All
	M (*SD*)	*M* (*SD*)	*M* (*SD*)	*M* (*SD*)	*M* (*SD*)	*M* (*SD*)
Correct responses	27.04 (6.32)	24.87 (6.04)	25.72 (6.23)	26.46 (5.13)	24.77 (4.79)	25.43 (4.99)
	Validity sample					
	Boys	Girls	All	Boys	Girls	All
	M (*SD*)	*M* (*SD*)	*M* (*SD*)	*M* (*SD*)	*M* (*SD*)	*M* (*SD*)
Correct responses	31.43 (6.41)	29.02 (6.25)	29.90 (6.37)	28.11 (8.65)	26.85 (6.55)	27.24 (7.13)

Concurrent Validity and Correlation With Reading Measures and General Cognitive Abilities

After using Holm correction to control for multiple comparisons, significant positive high correlations were found between the WRAT Math Computation scores and the three concurrent measures (range = .60-.80). The WRAT Math Computation subtest also correlated significantly with the descriptive measures, except with the reading comprehension scaled scores; these correlation strengths ranged from weak to moderate (range = .18-.52). Figure 4 shows the *r* values and the *p* values adjusted for multiple comparisons between the number of correct responses on the WRAT Math Computation subtests and all scaled scores of all the measures.

As mathematical abilities have been reported to correlate with other academic achievement tasks and cognitive abilities, we sought evidence that the correlation strengths between the WRAT Math Computation and other mathematical subtests were higher than those with subtests that evaluated nonmathematical abilities. We compared the largest *r* value of the correlation between the WRAT Math Computation subtest and one of the concurrent measures (ENI Written Math [speed]) with the largest *r* value of the correlation between that subtest and one of the descriptive measures (WISC IQ). The *r* value between the pair of measures (WRAT–ENI) was higher than that of the WRAT–IQ pair ($r = .52$, $Z = 3.75$, $p < .001$), providing evidence for adequate concurrent validity. Figure 5 shows a scatterplot of the WRAT Math Computation subtest scores and the speed scores from the ENI Written Math subtest.

Criterion Validity

We first defined the criteria to classify adolescents as having mathematical difficulty and mathematical talent. Adolescents were classified as having mathematical difficulty if they had either 17 or fewer correct responses on the WRAT Math Computation subtest or a score below the 10th percentile on the ENI Written Math subtest. Mathematical talent was defined as having either 34 or more correct responses on the WRAT Math Computation subtest or a score above the 90th percentile on the ENI Written Math subtest. As the ENI Written Math subtest has norm-referenced scores, its classification was used as a benchmark for comparing the classification of the WRAT Math Computation subtest.

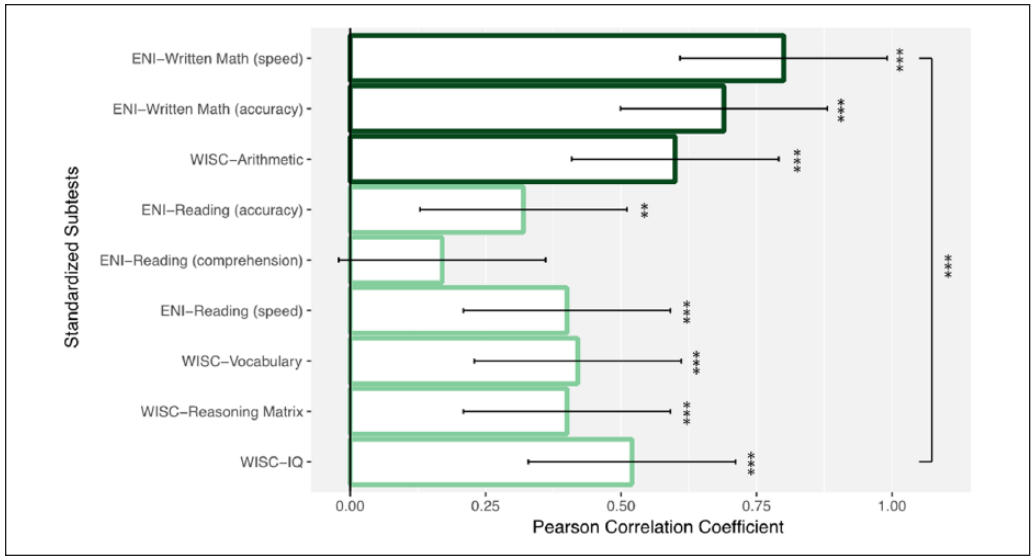

Figure 4. Pearson correlation coefficients between the number of correct responses and the scaled scores of the concurrent measures (dark green/black bars) and the scaled scores of the descriptive measures (light green/gray bars).
Note. Error bars represent 95% CI. ENI = Evaluación Neuropsicológica Infantil; WISC = Wechsler Intelligence Scale for Children; CI = confidence interval.
$^{**}p < .01$. $^{***}p < .001$.

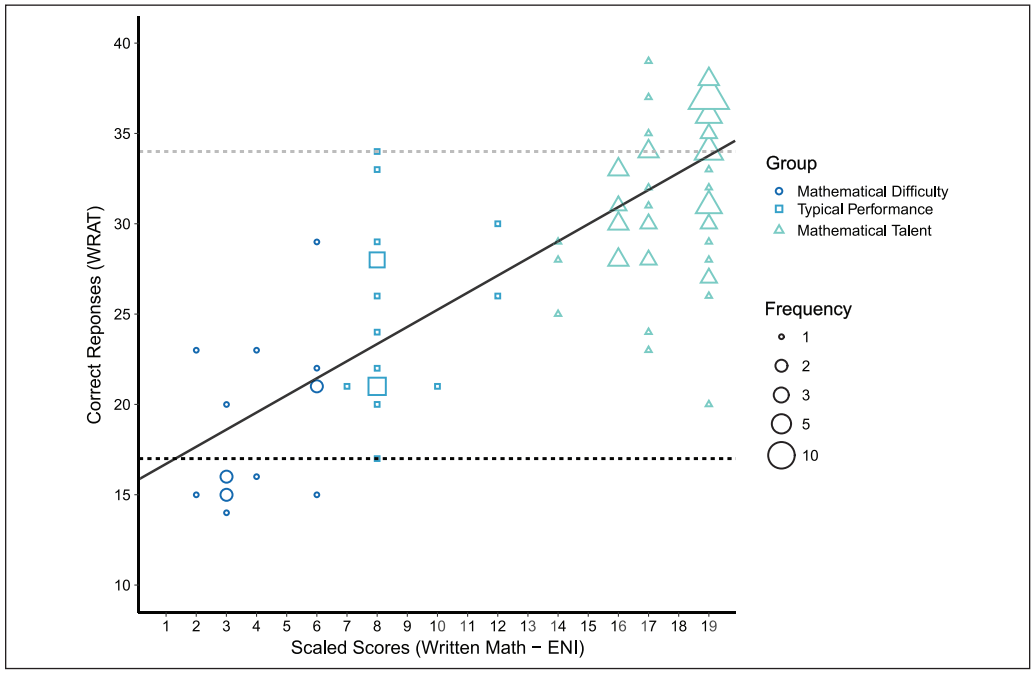

Figure 5. Scatterplot showing the association between the number of correct responses and the scaled scores of the ENI Written Math (speed) subtest.
Note. The figures represent the categorization of the participants according to the ENI. The dotted black line and the light gray line represent the cutoff scores according to the WRAT for mathematical difficulty and mathematical talent, respectively. ENI = Evaluación Neuropsicológica Infantil; WRAT = Wide Range Achievement Test.

Table 3. Cutoff Scores in Standard Deviations and Percentiles, Number of Cases of TP, FN, TN, FP, Sensitivity, and Specificity for Mathematical Difficulty and Mathematical Talent.

	Cutoff scores		TP	FN	TN	FP		
	WRAT	ENI	(+ +)	(− +)	(− −)	(+ −)	Sensitivity	Specificity
Mathematical difficulty	17 (–1.5 *SD*)	6 (<10th)	8	7	90	1	0.53	0.99
Mathematical talent	34 (+1.5 *SD*)	14 (>90th)	34	38	33	1	0.47	0.97

Note. First +/− sign indicates the classification according to the WRAT and the second that of the ENI. TP = true positives, FN = false negatives, TN = true negatives, FP = false positive, "+" = classified, "−" = not classified, sensitivity = TP / (TP + FN), specificity = TN / (TN + FP); WRAT = Wide Range Achievement Test; ENI = Evaluación Neuropsicológica Infantil.

Table 3 shows the number of adolescents classified with mathematical difficulty or mathematical talent by both subtests (true positives, TP), by the ENI Written Math subtests only (false negatives, FN), by the WRAT subtest only (false positives, FP), and those which were not classified by either subtest (true negatives, TN). It also shows the sensitivity and specificity of the WRAT Math Computation subtest relative to the ENI Written Math subtest. Although the sensitivity was low for mathematical difficulty and mathematical talent, the overall accuracy defined as $\frac{TP+TN}{N}$ (Ciccheti, 1994) for both categories was .92 and .78, respectively.

Psychometric Properties According to the Rasch Model

Unlike classical test theory (CTT), which fails to differentiate subject ability from item difficulty, the Rasch model (Rasch, 1960), a member of the item response theory family, measures participants' ability independent of the items administered. This tool is a one-parameter logistic model that defines an item difficulty as the position of that item on the latent dimension. Importantly, the Rasch model also gives us the opportunity to account for the unidimensionality of the subtest, as it considers all items to be part of the same ability. Using the TAM package of the statistical programming language R (Robitzsch, Kiefer, & Wu, 2018), we used a Rasch model to investigate the reliability, item fit, and internal structure of the WRAT Math Computation subtest.

The results were similar to those found with Cronbach's α: The expected a posteriori reliability also indicated good internal consistency (.86). Table 4 shows the item's infit statistics, where the expected value is one. According to Wilson (2005), items with an infit statistic outside the range 0.75 to 1.33, where the absolute value of the weighted *t* statistic is greater than 1.96, have a poor fit. As seen in Table 4, all 40 items showed a good fit to the Rasch model. For the sake of comparison, Table 4 also shows item difficulty according to the CTT (the percentage of participants who correctly answered that item) and according to the Rasch model.

The Wright map (Figure 6) provides a detailed graphical representation of how the Rasch model allows us to compare participant ability and item difficulty on the same scale. Higher logit values represent higher ability or greater difficulty. Adolescents' ability ranged from −4.77 to 5.28, whereas item difficulty ranged from −6.74 to 7.96, suggesting that the WRAT Math Computation subtest evaluates the entire range of their mathematical ability. A gap is visible, however, between Items 38 and 40, suggesting that Item 40, which requires reducing a rational expression to its lowest terms, might have been too difficult at this academic stage. Finally, the color of each number represents the factor it loaded to, according to EFA. Factor 2 items had greater estimated difficulty than those of Factors 1 and 3, consistent with the proportion of correct responses for each factor.

Table 4. Item Difficulty and Infit Statistics for Each Item.

Item	Difficulty (CTT approach)	Difficulty (Rasch model)	*SE*	Infit	*t*
1	99.24	−5.52	0.32	1.03	0.18
2	98.71	−4.98	0.25	1.01	0.12
3	99.77	−6.74	0.58	1.00	0.20
4	99.54	−6.04	0.41	1.01	0.15
5	98.10	−4.57	0.21	1.06	0.36
6	93.78	−3.27	0.12	1.16	1.66
7	96.74	−3.99	0.16	1.03	0.23
8	96.74	−3.99	0.16	1.04	0.32
9	91.81	−2.94	0.11	1.12	1.52
10	99.17	−5.43	0.31	0.99	0.05
11	93.17	−3.16	0.11	1.14	1.53
12	97.04	−4.10	0.17	0.97	−0.14
13	88.69	−2.54	0.09	1.04	0.66
14	85.58	−2.22	0.08	1.16	3.00
15	91.50	−2.90	0.10	1.05	0.62
16	79.36	−1.71	0.07	1.03	0.71
17	52.28	−0.14	0.06	1.10	3.68
18	67.22	−0.94	0.07	0.93	−2.65
19	59.48	−0.51	0.06	0.88	−4.73
20	67.53	−0.95	0.07	0.90	−3.48
21	84.90	−2.16	0.08	0.95	−1.02
22	83.08	−2.00	0.08	1.05	1.09
23	68.06	−0.98	0.07	0.94	−2.18
24	55.84	−0.32	0.06	1.00	−0.05
25	65.78	−0.86	0.07	1.16	5.23
26	64.80	−0.80	0.06	1.08	2.67
27	40.44	0.48	0.06	0.85	−5.61
28	28.83	1.15	0.07	1.13	3.84
29	45.98	0.19	0.06	0.88	−4.80
30	39.38	0.54	0.06	0.83	−6.66
31	18.44	1.88	0.08	0.91	−1.96
32	29.51	1.11	0.07	0.95	−1.68
33	41.12	0.45	0.06	1.04	1.41
34	34.37	0.82	0.07	0.90	−3.40
35	19.50	1.80	0.08	0.82	−4.43
36	41.50	0.43	0.06	1.10	3.61
37	22.69	1.56	0.07	0.89	−2.98
38	5.99	3.37	0.12	0.90	−1.08
39	11.91	2.50	0.09	0.91	−1.49
40	0.08	7.96	1.00	1.00	0.34

Note. Item difficulty according to the CTT is defined as the percentage of participants who correctly answered that item. CTT = classical test theory.

Discussion

This study demonstrates that the Math Computation subtest of the WRAT-IV (Wilkinson & Robertson, 2006) is a reliable and valid instrument to assess the mathematical abilities of 14- to

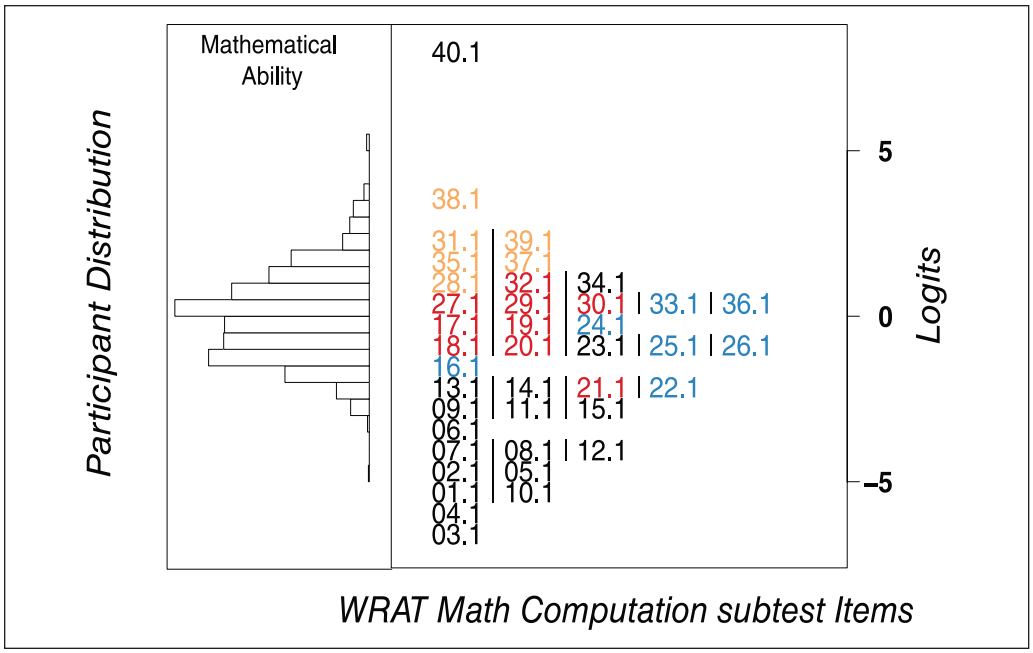

Figure 6. Wright map.
Note. The color of each number indicates the factor it loads to, according to the EFA: red = Factor 1, orange = Factor 2, and blue = Factor 3. Those in black did not load to any factor. EFA = exploratory factor analysis; WRAT = Wide Range Achievement Test.

16-year-old Mexican adolescents in their first year of high school, and it provides normative data that can be used to identify adolescents with mathematical difficulty and mathematical talent. We also report the underlying factor structure and the sex differences in the performance of this subtest.

First-year Mexican high school students had an average of 25.58 ($SD = 5.66$) correct responses on the WRAT Math Computation subtest. This mean is similar to that of 15- to 16-year-old adolescents in the United States (25.4, $SD = 8.0$), reported by Wilkinson and Robertson (2006). Nevertheless, this comparison is only an approximation, because we considered answers not written in their simplest form to be correct. Our Cronbach's α value of .86 and the expected a posteriori reliability indicate good internal consistency for the overall subtest. Our results also showed that the WRAT Math Computation subtest has adequate concurrent validity with two individually administered assessments of mathematical abilities (the ENI Written Math and WISC Arithmetic subtests). There was also agreement between the scores of the WRAT Math Computation subtest and the descriptive measures (e.g., the WISC IQ); however, the r value of the correlation between the Math Computation and the Written Math subtests was higher than the correlation with any other descriptive measure.

To explore the psychometric properties of the WRAT math computation subtest, we used both EFA and Rasch models. Although these approaches are based on different assumptions, they gave us the opportunity to determine the test's underlying factors and measure a person's mathematical ability at the item level. Results from the EFA suggested a three-factor model for the adolescents in this study. These factors were labeled fractions and basic algebra, rational numbers, and arithmetic. Items that loaded on these three factors tell us about the arithmetical abilities participants are still learning, as the proportion of correct responses for none of these factors was higher than .62. In general, performance on these items shows that adolescents are still in the

process of learning how to solve arithmetical operations with rational numbers and that they have particular difficulty when the relationship between rational and whole number arithmetic is more opaque. This was the case for the questions in the rational number factor. Problems in this factor involved solving arithmetic problems involving fractions with unlike denominators and converting decimal numbers with leading zeros to percentages or fractions. These types of rational number problems have been proposed as inherent sources of difficulty (Siegler & Lortie-Forgues, 2017). There is, thus, a need to reinforce the understanding of rational numbers in high school, as it is an important predictor of performance in algebra (Siegler et al., 2012).

The three-factor model also suggests that adolescents' math abilities at this academic stage may not be global ability; instead, they might be made up of their different proficiency levels in solving arithmetic problems with rational numbers (i.e., all numbers that can be expressed as a/b) with different formats (e.g., decimals, fractions, and percentages). The proportion of correct responses for these factors shows that an item's difficulty depends on the rational number format. Although it is still a controversial subject (Tian & Siegler, 2018), decimals and fractions may have different intrinsic difficulties.

Sex differences in mathematical achievement are highly dependent on the type of measure used and the developmental stage in which they are evaluated. Whereas most studies find higher performance in boys (Diamantopoulou, Pina, Valero-Garcia, González-Salinas, & Fuentes, 2012; Stoet & Geary, 2015), others have found that girls outperform boys in curriculum-based measurements (Yarbrough, Cannon, Bergman, Kidder-Ashley, & McCane-Bowling, 2017), and some find no sex differences in the performance of early numerical tasks (Beltrán-Navarro, Abreu-Mendoza, Matute, & Rosselli, 2016; Rosselli et al., 2009). Our results show that Mexican male students in the first year of high school outperform their female peers on the WRAT Math Computation subtest. On average, male students had two more correct responses. However, the size effect of these differences was small (Cohen's $d = 0.34$), with an 87% overlap between the two groups. Future research is needed to understand the implications of these differences.

Our results also showed that the WRAT Math Computation subtest is a useful screening assessment to identify adolescents with mathematical difficulty or mathematical talent. Almost all participants meeting these criteria received similar classifications using the ENI Written Math subtest. However, a number were not classified with mathematical difficulty or talent based on their ENI score, but were so classified using the WRAT score. Lenient cutoff scores for the WRAT Math computation subtest could reduce the number of FN; however, a cost–benefit analysis should be considered to avoid FP.

Psychometric properties from the Rasch model indicated that the WRAT Math Computation subtest has an appropriate internal structure and that its items appropriately assess the mathematical ability of first-year Mexican high school students. Analysis of the difficulty of different items confirmed that low performance on items that involved problems with rational numbers was not necessarily an effect of the order in which they appeared. Item 31, for example, which involved changing a decimal number to a percentage, was more difficult than Item 36, which involved a multidigit whole number division.

The limitations of this study are several. Although the sociodemographic backgrounds of our participants, as measured by maternal education, were diverse, all participants attended schools in the same Mexican city, so the results may not be generalizable to other regions of Mexico. However, the agreement between the WRAT Math Computation scores and Mexican standardized assessments suggests that this is not the case. Another limitation was the low internal reliability of Factor 3, which could be the consequence of mixing whole and decimal numbers. Although the number of digits was similar across items in this factor, some numbers included decimals. More important, the eigenvalue to justify a third factor was low, so conclusions based on the scores for this factor should be made with caution. Still, Factor 3 could be useful for evaluating adolescents' ability to solve multidigit arithmetical problems and contrasting it with their

ability to solve arithmetical problems including fractions (Factors 1 and 2). A final limitation is the low sensitivity of the subtest. As previously suggested, using lenient cutoff scores could, in some cases, ameliorate this problem. If educational practitioners' aim is to detect adolescents with mathematical difficulty, FP would have less impact than if the aim were to characterize their cognitive profiles.

In summary, our study provides reliable and valid normative data for the WRAT Math Computation subtest for Mexican adolescents in the first year of high school. Cutoff points based on the number of correct responses in this subtest correctly identify adolescents with mathematical difficulty and mathematical talent. Finally, the underlying factors found using the WRAT Math Computation subtest could help to develop a more finely detailed description of mathematical ability during adolescence.

Authors' Note

Roberto A. Abreu-Mendoza is now affiliated with Rutgers University, USA.

Acknowledgments

The authors wish to thank Diana Ávalos and Daniel Romero for their help in evaluation and data collection, and Gwennaëlle Pupier and Hillary Contreras for their assistance in database collection. We also thank all participants and their families, and the high schools, principals, and teachers who assisted with this project.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This work was supported by two grants from the *Consejo Nacional de Ciencia y Tecnología* (CONACyT): Programa Nacional de Posgrado de Calidad Fellowship No. 576221 (awarded to the first author) and Investigación en Fronteras de la Ciencia Grant No. 783 (awarded to the corresponding author).

Note

1. The six-factor model was also tested and had a good fit; however, this model had two items that split across two factors and two factors that each had only one high loading. For these reasons and for the sake of structural simplicity, the three-factor model was selected.

Supplementary Material

The complete data set and analysis codes can be found on the Open Science Framework project page: https://osf.io/xaznt/

ORCID iD

Roberto A. Abreu-Mendoza https://orcid.org/0000-0002-6841-0917

References

American Educational Research Association, National Council on Measurement in Education, & American Psychological Association. (2014). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

Beltrán-Navarro, B., Abreu-Mendoza, R. A., Matute, E., & Rosselli, M. (2016). Development of early numerical abilities of Spanish-speaking Mexican preschoolers: A new assessment tool. *Applied Neuropsychology: Child*, *7*, 117-128. doi:10.1080/21622965.2016.1266940

Buelow, M., & Frakey, L. L. (2013). Math anxiety differentially affects WAIS-IV arithmetic performance in undergraduates. *Archives of Clinical Neuropsychology*, *28*, 356-362. doi:10.1093/arclin/act006

Catell, R. B. (1966). The scree test for the number of factors. *Multivariate Behavioral Research*, *1*, 245-276.

Ciccheti, D. (1994). Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. *Psychological Assessment*, *6*, 284-290.

Diamantopoulou, S., Pina, V., Valero-Garcia, A. V., González-Salinas, C., & Fuentes, L. J. (2012). Validation of the Spanish version of the Woodcock-Johnson mathematics achievement tests for children aged 6 to 13. *Journal of Psychoeducational Assessment*, *30*, 466-477. doi:10.1177/0734282912437531

Hoff, E. (2006). How social contexts support and shape language development. *Developmental Review*, *26*, 55-88.

Horn, J. L. (1965). A rationale and test for the number of factors in factor analysis. *Psychometrika*, *30*, 179-185.

Hyde, J. S., Fennema, E., & Lamon, S. (1990). Gender differences in mathematics performance: A meta-analysis. *Psychological Bulletin*, *107*, 139-155.

Instituto Nacional para la Evaluación de la Educación (INEE). (2015). *Panorama Educativo de México 2014. Indicadores del Sistema Educativo Nacional* [Educational Panorama of Mexic 2014. Indicators of the National Educational System]. Educación Básica y Media Superior. México: Author.

Instituto Nacional de Estadística y Geografía. (2015). *Panorama sociodemográfico de México 2015* [Sociodemographic Panorama of Mexico 2015]. Ciudad de Mexico, México: INEGI.

Jolliffe, I. T. (1986). *Principal component analysis*. New York, NY: Springer.

Matute, E., Rosselli, M., Ardila, A., & Ostrosky, F. (2007). *Evaluación Neuropsicológica Infantil* [Neuropsychological Assessment of Children]. México: Manual Moderno.

McCauley, R., & Swisher, L. (1984). Psychometric review of language and articulation tests for preschool children. *Journal of Speech and Hearing Disorders*, *49*, 34-42.

Organisation for Economic Co-operation and Development. (2014). *PISA 2012 results: What students know and can do—Student performance in mathematics, reading and science* (Vol. 1, Rev. ed.). Paris: OECD Publishing.

Preacher, K. J., & MacCallum, R. C. (2003). Repairing Tom Swift's electric factor analysis machine. *Understanding Statistics*, *2*, 13-32.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Chicago, IL: The University Chicago Press.

Revelle, W. (2017). *psych: Procedures for personality and psychological research*. Evanston, IL: Northwestern University. Retrieved from https://cran.r-project.org/package=psych

Ritchie, S. J., & Bates, T. C. (2013). Enduring links from childhood mathematics and reading achievement to adult socioeconomic status. *Psychological Science*, *24*, 1301-1308.

Robitzsch, A., Kiefer, T., & Wu, M. (2018). *TAM: Test analysis modules*. Kiel, Germany: Kiel University. Retrieved from https://github.com/alexanderrobitzsch/TAM

Rosselli, M., Ardila, A., Matute, E., & Inozemtseva, O. (2009). Gender differences and cognitive correlates of mathematical skills in school-aged children. *Child Neuropsychology*, *15*, 216-231.

Rosselli, M., Matute, E., Pinto, N., & Ardila, A. (2006). Memory abilities in children with subtypes of dyscalculia. *Developmental Neuropsychology*, *30*, 801-818.

Sattler, J. (2010). *Evaluacion infantil: Fundamentos cognitivos* [Assessment of children: Cognitive foundations] (Vol. I, 5th ed.). Ciudad de Mexico, Mexico: Manual Moderno.

Schmitt, T. A. (2011). Current methodological considerations in exploratory and confirmatory factor analysis. *Journal of Psychoeducational Assessment*, *29*, 304-321.

Siegler, R. S., Duncan, G. J., Davis-Kean, P. E., Duckworth, K., Claessens, A., Engel, M., . . . Chen, M. (2012). Early predictors of high school mathematics achievement. *Psychological Science*, *23*, 691-697. doi:10.1177/0956797612440101

Siegler, R. S., & Lortie-Forgues, H. (2017). Hard lessons: Why rational number arithmetic is so difficult for so many people. *Current Directions in Psychological Science*, *26*, 346-351. doi:10.1177/0963721417700129

Starr, A., DeWind, N., & Brannon, E. M. (2017). The contributions of numerical acuity and non-numerical stimulus features to the development of the number sense and symbolic math achievement. *Cognition*, *168*, 222-233.

Stoet, G., & Geary, D. C. (2015). Sex differences in academic achievement are not related to political, economic, or social equality. *Intelligence*, *48*, 137-151.

Sturges, H. A. (1926). The choice of a class interval. *Journal of the American Statistical Association*, *21*, 65-66.

Tian, J., & Siegler, R. S. (2018). Which type of rational numbers should students learn first? *Educational Psychology Review*, *30*, 351-372.

Wechsler, D. (2007). *Escala Wechsler de Inteligencia para Niños-IV* [Wechsler Intelligence Scale for Children-IV]. Ciudad de Mexico, Mexico: Manual Moderno.

Wilkinson, G. S., & Robertson, G. J. (2006). *WRAT 4 Wide Range Achievement Test*. Lutz, FL: Psychological Assessment Resources.

Wilson, M. (2005). *Constructing measures: An item response modeling approach*. Mahwah, NJ: Lawrence Erlbaum.

Yarbrough, J. L., Cannon, L., Bergman, S., Kidder-Ashley, P., & McCane-Bowling, S. (2017). Let the data speak: Gender differences in math curriculum–based measurement. *Journal of Psychoeducational Assessment*, *35*, 568-580. doi:10.1177/0734282916649122