Validation of the Spanish Version of the Woodcock-Johnson Mathematics Achievement Tests for Children Aged 6 to 13

Sofia Diamantopoulou, Violeta Pina, Ana V. Valero-Garcia, Carmen González-Salinas and Luis J. Fuentes

Journal of Psychoeducational Assessment 2012 30: 466 originally published online 4 April 2012
DOI: 10.1177/0734282912437531

Journal of Psychoeducational Assessment
30(5) 466–477

DOI: 10.1177/0734282912437531
http://jpa.sagepub.com

Validation of the Spanish Version of the Woodcock-Johnson Mathematics Achievement Tests for Children Aged 6 to 13

Sofia Diamantopoulou[1], Violeta Pina[1], Ana V. Valero-Garcia[1], Carmen González-Salinas[1], and Luis J. Fuentes[1]

Abstract

This study validated the four mathematics tests of the Spanish version of the Woodcock-Johnson III (WJ-III) Achievement (ACH) battery for use in the first six grades of school in Spain. Developmental effects and gender differences were also examined. Participants were a normal population sample of 424 (216 boys) children aged 6 to 13 years. Results showed that the tests have good test-retest and internal reliability and good construct and criterion-related validity. Significant main effects of schooling were obtained with scores increasing across the six school grades, but scores between fourth and fifth graders did not differ significantly. Overall, boys scored higher than girls on all tests but the effect sizes of these gender differences were small ($d \leq .12$).

Keywords

mathematical ability, psychometric validation, gender differences, age effects, Batería III APROV

Introduction

Approximately 7% of children and adolescents experience severe difficulties in at least one area of mathematics (Barbaresi, Katusic, Colligan, Weaver, & Jacobsen, 2005; Shalev, Manor, & Gross-Tsur, 2005). Individuals' performance with mathematical learning disability (MLD) on mathematic achievement tests is substantially below an expectation based on chronological age, general intellectual ability, and age-appropriate educational history (American Psychiatric Association, 2000). A comprehensive instrument for the evaluation of academic achievement is the Woodcock-Johnson III (WJ-III) Achievement (ACH) battery (Woodcock, McGrew, & Mather, 2001, 2007). The four mathematics tests of the WJ-III ACH battery include speeded and nonspeeded tests, tests of simple and complex mathematic problems that require or not the use

[1]University of Murcia, Murcia, Spain

Corresponding Author:
Sofia Diamantopoulou, Department of Basic Psychology and Methodology, University of Murcia, Campus Espinardo, 30100 Murcia, Spain
Email: sofiad@um.es

of language and memory skills, and finally, tests of knowledge of core mathematical concepts, symbols, and vocabulary (Woodcock et al., 2001, 2007).

The WJ-III ACH battery has been translated to Spanish (Batería III Woodcock-Muñoz, Pruebas de aprovechamiento [Batería III APROV]; Muñoz-Sandoval, Woodcock, McGrew, & Mather, 2005) and normed on samples representing several regions of the Spanish-speaking world. The Spanish version of the WJ-III ACH battery has shown good psychometric properties in independent studies and the four math tests have been found to correlate significantly with school grades in mathematics (e.g., Rodriguez & Diaz, 1990; Rosselli, Ardila, Bateman, & Guzmán, 2001). However, a closer look at the manual suggests that caution may be in order for its use in Spain given that the Spanish calibration only included 10 subjects from Spain. To date, no independent studies have examined the validity of the Batería III APROV in Spain. Because cultural factors play a central role in achievement test performance (Ardila, 1995), current norms of the Batería III APROV may be misleading when used in Spain.

The few existing Spanish studies of children's mathematic ability have relied on nonstandardized achievement tests that did not measure mathematic ability on a common scale (e.g., Alsina & Sáiz, 2003; Martín, Martínez-Arias, Marchesi, & Pérez, 2008), or they used the Arithmetic ability test of the Wechsler Intelligence Test for Children (WISC; Miranda, Meliá, Marco, Roselló, & Mulas, 2006), or finally, they relied on school grades (e.g., Checa, Rodriguez-Bailon, & Rueda, 2008). Perhaps the most important results about academic achievement in Spanish samples come from two international studies: the Programme for International Student Assessment study (PISA) and the Trends in International Mathematics and Science Study (TIMSS). The results of both the above studies place the average mathematics performance of Spanish children in the lowest rank and below the performance of their American peers (Organisation for Economic Cooperation and Development [OECD], 2009; Peak, 1996). However, the age of the participants in the PISA study has been 15 to 16 years, whereas only children in the fourth and eighth school grades are assessed in the TIMSS. Because the most rapid academic growth in mathematics is observed in the first three school grades (Lee, 2010; Lichten, 2004; McGrew, Schrank, & Woodcock, 2007), a question that remains unanswered is whether Spanish students follow, or not, the same developmental trend in mathematics as reported in the international literature.

This study is the first to validate the math-achievement tests of the Batería III APROV for use in the first six grades of school in Spain. We aimed to examine whether the tests are appropriate for evaluating mathematic ability across primary education by significantly differentiating ability across the first six school grades of primary education. We also aimed to examine whether the development of Spanish students' mathematic ability follows the same developmental trend reported in international studies. Finally, because a recent meta-analysis of the PISA and the TIMSS findings showed a small gender gap in mathematics achievement in Spain with adolescent boys performing better than girls (Else-Quest, Hyde, & Linn, 2010), we examined whether this would be observed also in the first school years.

Method

Participants

A total of 424 children (216 boys) aged 6 to 13 years (M = 9 years, SD = 2 months) and their parents participated in this study that was conducted in the Region of Murcia, in South-Eastern Spain. Descriptive data are shown in Table 1. To obtain a sample representative of the normal population of children, we selected 8 urban and suburban schools in low-, medium-, and high socioeconomic status residential areas. We selected a random sample of approximately 8 children

Table 1. Descriptive Data

		Age (months)		
Grade	*n* (boys)	*M*	*SD*	Range
1	82 (45)	82.4	3.6	74-89
2	64 (32)	93.7	3.9	86-101
3	64 (30)	106.5	4.1	98-118
4	66 (33)	117.6	5.1	108-132
5	75 (37)	130.6	5.0	121-147
6	73 (39)	143.7	4.9	135-158
Total	424 (216)			

(50% boys) per grade and school to administer the tests. Parents completed a questionnaire pertaining to demographic characteristics. Parental questionnaires were obtained for 387 children after two reminders.

All children spoke fluent Spanish but, at home, 4 children spoke Arabic, 6 children spoke some other European language (e.g., Russian, Ukrainian), and 19 children spoke both Spanish and a second language (e.g., both Spanish and Arabic). The majority of children (74%) were born in Spain by Spanish parents. Fourteen percent of the children were born in Spain by one or both non-Spanish parents and 8% were born outside Spain by non-Spanish parents. For the remaining children, their or their parents' nationality is unknown.

The majority of children (87%) lived with both biological parents, whereas 13% of the children lived with only one biological parent. Family income reports ($N = 329$) in euros per month were as follows: 11% reported less than 750, 19% reported between 751 and 1,200, 17% reported between 1,201 and 1,600, 10% reported between 1,601 and 2,000, 14% reported between 2,001 and 3,000, 15% reported more than 3,000, and for the remaining families monthly income is unknown. Parental education, based on data from 378 mothers and 360 fathers, was as follows: 10% of the mothers and 13% of the fathers reported 6 years of schooling, 63 % of mothers and 60% of fathers reported 12 years of schooling, and the remaining had a college or university degree.

Procedure

The study was approved by the bioethics committee of the University of Murcia. Tests were administered individually by trained assistants in one session lasting, depending on the child's age, from 30 to 45 min. We obtained written informed consent from all parents and children gave their oral informed consent at the beginning of the testing. Tests were administered in a counterbalanced sequence to eliminate systematic variations due to order of administration. Counterbalancing was achieved by randomly ordering protocols for the tests for each child within each class using a table of random digits.

Native Spanish speakers adapted some instructions of the tests Applied problems and Quantitative Concepts so that they would better correspond with the Spanish spoken in Spain (see appendix). The changes were minimal and did not change the content of the items.

To calculate test-retest reliability for the Batería III APROV, 271 children (141 boys) were administered the test twice within a 1-week interval. We considered a 1-week test-retest time interval sufficient for ruling out memory effects and short enough to rule out learning effects as children are introduced to new mathematic concepts weekly.

Measures

Math ability tests of the Batería III APROV. The Batería III APROV is designed for use with subjects from preschool (from age 2 years) to geriatric levels. The four tests of the battery that assess mathematical ability are Calculation, Math Fluency, Applied Problems, and Quantitative Concepts. The scoring for all items is 0 (*incorrect*) and 1 (*correct*).

Calculation measures the ability to perform simple mathematical computations including addition, subtraction, multiplication, and division. The test contains 45 problems of ascending difficulty presented in a subject response booklet. The test has no time limit but no scores are counted for items solved after six consecutive mistakes.

Math Fluency measures the ability to solve simple addition, subtraction, and multiplication facts quickly. The test consists of a series of 160 simple arithmetic problems in a subject response booklet and the child needs to complete as many as possible in a 3-min time limit.

Applied Problems measures the ability to analyze and solve math problems. To solve the problems, the child is required to listen to the problem, recognize the procedure to be followed, and then perform relatively simple calculations. Because many of the problems include extraneous information, the child needs to decide not only the appropriate mathematical operations to use but also what information to include in the calculation. The test consists of 62 problems of ascending difficulty that are read to the child and testing stops once the child has made six consecutive mistakes.

Quantitative Concepts measures knowledge of mathematical concepts, symbols, and vocabulary. There are two subtests administered: Concepts and Number Series. In the first subtest, the child is required to count, use numbers and concepts, and identify mathematical terms. This test consists of 34 items of ascending difficulty that are read to the child and testing stops once the child has made four consecutive mistakes. In the second subtest, the task requires the child to look at a series of numbers, figure out the pattern, and then provide the missing number in the series. This test consists of 23 problems of ascending difficulty and testing stops once the child has made three consecutive mistakes.

Intelligence. We assessed *nonverbal intelligence* with the Spanish version of the Matrices subtest of the Kaufman Brief Intelligence Test (K-BIT; Kaufman & Kaufman, 1990). The validity of the Matrices subtest of the Spanish version of the K-BIT measured as coefficient alpha is excellent ranging from $\alpha = .86$ to $\alpha = .93$ respectively for the ages 6 to 12 years (Kaufman & Kaufman, 1990). The relation between nonverbal intelligence and mathematics achievement is well established in the literature (e.g., Alarcón, Knopik, & DeFries, 2000).

Arithmetic Ability (WISC). We assessed arithmetic ability with the arithmetic ability test of the Spanish version of the Wechsler Intelligence Test for Children (WISC-IV; 2004). Internal consistency for the Spanish version of this test is excellent with alphas ranging from $\alpha = .82$ to $\alpha = .88$ for the ages 6 to 12, respectively, whereas a recent clinical study has provided support for the criterion validity of the Spanish WISC-IV tests (Moines, Allen, Puente, & Neblina, 2010).

Statistical Analyses

The construct validity of the math ability tests of the Batería III APROV was evaluated using confirmatory factor analyses (CFA) with data from the first assessment of the 424 children. This assessed the dimensionality of the tests confirming that all items measured the same common construct, permitting unbiased scaling of individuals on a common ability (Wirth & Edwards, 2007). In line with the statistical analysis plan of the test manual of the Batería III APROV, construct validity (and internal consistency) of the test Fluency was not evaluated as it is a speeded test that consists of problems of equal difficulty.

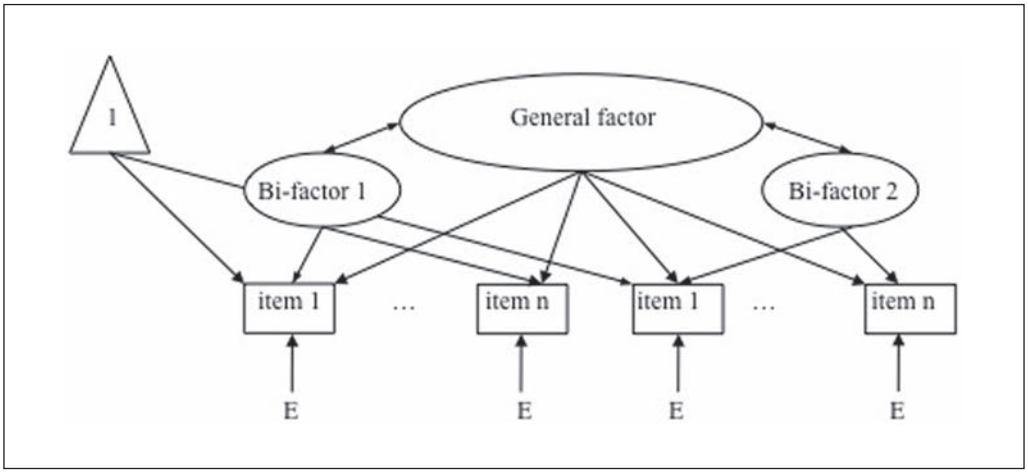

Figure 1. Illustration of a bifactor confirmatory factor analysis mode

We first fitted CFA models to the data where all items loaded on one common factor. To account for the dichotomous item responses, we applied Weighted Least Squares Estimation (WLSE) analyses to tetrachoric correlations between item scores. If the fit of the one-factor model was poor, we conducted exploratory factor analyses (EFA), the results of which guided us for conceptualizing bifactor CFA models. Bifactor models are defined by a single "general" factor that accounts for the common variance among all items, and method factors that take into account additional systematic variability that remains between the items after accounting for the shared variance with the general factor (see Figure 1). The fit of the models was evaluated according to the guidelines outlined by Hu and Bentler (1999): comparative fit index CFI ≥ .95, Tucker-Lewis index TLI ≥ .95, and root mean square error of approximation RMSEA ≤ .06. However, it has been shown that CFI and TLI are often more dependable than other fit measures, whereas a strict RMSEA ≤ .06 limit has been found to overreject theoretically sound models (Yu, 2002). Therefore, we accepted models that provided good fit statistics according to the CFI and TLI even if RMSEA > .06. All further analyses included the items analyzed in the final CFA models. All models were estimated using Mplus version 5 (Muthén & Muthén, 1998-2007).

Having established the dimensionality of each test, we assessed the internal consistency of the tests by calculating coefficient alphas. We next examined school grade and sex effects on test performance with analyses of variance (ANOVA). We assessed the criterion-related validity of the four-math ability tests of the Batería III APROV by examining bivariate correlations between children's scores on the four tests and their scores on the WISC-arithmetic and nonverbal IQ. Finally, we assessed test-retest reliability by correlating children's results on the four tests of the Batería III APROV at Time 1 and Time 2.

Results

Construct validity. Results of the CFA for the test Calculation revealed that a one-factor model including 27 items, that is, Items 5 to 32 of the test, excluding Items 1 to 4 that were correctly answered by all children and Items 26 and 33 to 45 that no children answered correctly, did not fit the data well: $\chi^2(53) = 445.95$, $p < .01$, CFI = .89, TLI = .93, RMSEA = .13. Results of the EFA revealed that a two-factor model, where Items 5 to 22 loaded on the first bifactor and Items 23 to 32 loaded on the second bifactor, fitted the data best: $\chi^2(298) = 713.83$, $p < .01$, CFI = .98, TLI = .98,

RMSEA = .05. The bifactor CFA model we fitted based on the results of the EFA provided a good fit to the data: $\chi^2(50) = 270.23$, $p < .01$, CFI = .95, TLI = .96, RMSEA = .10. Items loading on the second bifactor differed from items loading on the first bifactor in terms of difficulty and content as they included apart from addition, subtraction, multiplication, and division, as items in Factor 1, also subtraction and division of fractions with different denominators, multiplication of numbers containing decimals, and estimation of percentages. As seen in Table 2, all items loaded significantly on the common factor in the bifactor CFA.

Results for the test Applied Problems revealed that a one-factor CFA model including 40 items, that is, Items 7, and 9 to 47 (remaining items were excluded due to zero variance, i.e., all children answered Items 1-6 and Item 8 correctly and no children answered Items 48 to 62 correctly) provided a good fit to the data $\chi^2(30) = 122.24$, $p < .01$, CFI = .95, TLI = .95, RMSEA = .08. All items loaded significantly on the common factor in the bifactor CFA (see Table 2).

For Quantitative Concepts a one-factor model including 38 items, that is, Items 8 to 29 of the subtest Concepts and Items 7 to 22 of the subtest Series (remaining items were excluded because of zero variance, i.e., all children answered correctly Items 1 to 7 of the test Concepts and Items 1 to 6 of the test Series and no children answered the remaining items correctly) did not provide a good fit to the data: $\chi^2(37) = 607.35$, $p < .01$, CFI = .90, TLI = .94, RMSEA = .12. Results of the EFA showed that a two-factor model, where Items 8 to 25 of the subtest Concepts and Items 7 to 17 of the subtest Series loaded on the first bifactor, whereas remaining items loaded on the second bifactor, provided a good fit to the data: $\chi^2(628) = 2{,}458.23$, $p < .01$, CFI = .98, TLI = .97, RMSEA = .08. Items loading on the second bifactor differed from items loading on the first bifactor in terms of content and difficulty. As regards Items 18 to 22 for the subtest Concepts, these items included addition of fractions with different denominators and equation, that is, calculations that are not included in Items 1 to 17 in the test that mainly require simple calculations of addition and subtraction. As regards the test Series, whereas items up to 17 require a simple addition or subtraction to identify the correct answer (e.g., Item 11: _ 3 5 7, correct answer is 1), Items 18 to 22 require multiplication and division to identify the correct answer (e.g., Item 19: 81 27 9 _, correct answer is 3). The bifactor CFA model we fitted based on the results of the EFA provided a good fit to the data: $\chi^2(72) = 323.92$, $p < .01$, CFI = .95, TLI = .96, RMSEA = .09. All items loaded significantly on the common general factor in the bifactor CFA (see Table 2).

Internal consistency. The internal consistency of the tests was excellent: Calculation α =.90, Applied Problems α = .91, and Quantitative Concepts α = .92.

Schooling and sex effects. Mean values by school grade and sex are depicted in Table 3 as well as effect sizes (i.e., Cohen's *d*) calculated for the significant differences in performance between consecutive grades.

We obtained significant effects of school grade for Calculation, $F(5, 424) = 212.11$, $p < .01$, Fluency, $F(5, 424) = 126.24$, $p < .01$, Applied Problems, $F(5, 424) = 157.10$, $p < .01$, and Quantitative Concepts, $F(5, 424) = 148.58$, $p < .01$, indicating that math ability increased with years of schooling. However, Tukey´s post hoc tests revealed that children in Grades 4 and 5 did not differ significantly in their performance in any of the four math tests of the Batería III APROV. As shown in Table 3, differences in performance, and consequently effect sizes, were larger for the first three school grades except for the test Calculation where the score difference between Grades 5 and 6 was larger than the score difference between Grades 3 and 4.

We obtained significant effects of sex for the tests Applied Problems, $F(1, 424) = 5.93$, $p < .05$, for Fluency, $F(1, 424) = 4.86$, $p < .05$, and for Quantitative Concepts, $F(1, 424) = 6.34$, $p < .01$, indicating that boys, overall, performed better than girls. The effect size of these sex differences calculated as Cohen's *d* was small: Fluency $d = .12$, Applied Problems $d = .11$, Quantitative Concepts $d = .11$. We obtained no significant interaction effects of sex and school grade (Fs ≤ 1.06, *ns*).

Table 2. Standardized Factor Loadings Based on Confirmatory Factor Analyses

	Test			
			Quantitative Concepts	
Item	Calculation	Applied Problems	Concepts	Series
5	.21			
6	.28			
7	.29	.46		.62
8	.32		.58	.72
9	.54	.43	.19	.57
10	.65	.30	.43	.80
11	.85	.37	.32	.73
12	.55	.66	.63	.78
13	.92	.50	.43	.74
14	.74	.53	.80	.81
15	.95	.38	.91	.33
16	.83	.68	.94	.70
17	.91	.60	.87	.77
18	.94	.64	.97	.53
19	.88	.65	.87	.54
20	.82	.71	.91	.46
21	.88	.76	.83	.33
22	.86	.80	.84	.30*
23	.75	.88	.79	
24	.58	.75	.92	
25	.88	.76	.71	
26		.75	.56	
27	.62	.79	.62	
28	.78	.83	.64	
29	.60	.77	.48	
30	.63	.72		
31	.78	.90		
32	.83	.81		
33		.86		
34		.88		
35		.77		
36		.72		
37		.81		
38		.74		
39		.69		
40		.70		
41		.72		
42		.76		
43		.72		
44		.61		
45		.69		
46		.70		
47		.38		

Note: All factor loading significant at $p < .01$ level except for loadings marked with * that were significant at $p < .05$ level. Omitted items had zero variation.

Table 3. Mean Total Scores by Grade and Sex and Effect Sizes (D) for Significant Differences in Scores Between Consecutive School Grades

	Calculation			Fluency			Applied Problems			Quantitative Concepts		
	All	Boys/girls		All	Boys/girls		All	Boys/girls		All	Boys/girls	
	M	*M*		*M*	*M*		*M*	*M*		*M*	*M*	
Grade	*SD*	*SD*	*d*	*SD*	*SD*	*d*	*SD*	*SD*	*d*	*SD*	*SD*	*d*
1	5.68	5.75/5.59		26.28	27.71/24.54		17.25	17.62/16.81		9.02	9.15/8.86	
	2.07	1.83/2.36		9.93	9.33/10.48		3.80	3.71/3.90		4.18	3.41/5.00	
2	9.79	9.93/9.65	2 vs. 1	37.48	38.38/36.59	2 vs. 1	21.34	22.25/20.43	2 vs. 1	14.06	15.12/13.00	2 vs. 1
	2.79	2.86/2.76	1.67	9.76	8.36/11.02	1.13	3.78	3.91/3.47	1.07	4.69	4.64/4.56	1.13
3	13.39	13.16/13.58	3 vs. 2	49.41	51.23/47.79	3 vs. 2	25.29	25.23/25.35	3 vs. 2	18.23	18.30/18.17	3 vs. 2
	2.54	3.32/2.74	1.34	13.74	14.73/12.04	1.00	4.74	4.74/4.81	0.92	4.68	4.14/5.18	0.89
4	14.90	15.27/14.54	4 vs. 3	59.86	62.42/57.30	4 vs. 3	28.95	29.39/28.51	4 vs. 3	21.25	22.24/20.27	4 vs. 3
	1.98	1.50/2.33	0.66	13.41	14.39/12.04	0.76	3.91	3.77/4.05	0.84	4.00	3.49/4.28	0.69
5	15.41	15.70/15.13	5 vs. 4	64.31	67.81/60.89	5 vs. 4	30.18	30.40/29.97	5 vs. 4	22.68	22.56/22.78	5 vs. 4
	2.79	2.45/3.09	NA	16.93	18.61/14.56	NA	4.09	3.80/4.39	NA	4.02	4.74/3.22	NA
6	18.13	18.53/17.67	6 vs. 5	76.36	75.38/77.47	6 vs. 5	32.15	33.00/31.17	6 vs. 5	24.86	25.76/23.82	6 vs. 5
	3.52	3.62/3.29	0.85	19.00	20.65/17.15	0.66	3.41	3.02/3.62	0.52	3.94	4.46/2.98	0.54

Table 4. Correlations Between the Mathematic Achievement Tests and Intelligence at Test (1) and Retest (2) Occasions

	1.	2.	3.	4.	5.	6.	7.	8.	9.	10.
1. Calculation (1)	1									
2. Fluency (1)	.81	1								
3. Applied Problems (1)	.82	.80	1							
4. Quantitative Concepts (1)	.83	.78	.88	1						
5. Calculation (2)	.93	.83	.83	.85	1					
6. Fluency (2)	.81	.95	.77	.76	.81	1				
7. Applied Problems (2)	.86	.81	.94	.88	.86	.79	1			
8. Quantitative Concepts (2)	.85	.80	.86	.95	.87	.77	.88	1		
9. WISC-arithmetic (1)	.78	.75	.84	.82	.78	.72	.84	.82	1	
10. Nonverbal intelligence (1)	.65	.58	.69	.70	.70	.62	.71	.71	.67	1

Note: All correlations significant at $p < .01$ level.

Criterion-related validity. As shown in Table 4, all four mathematical achievement tests of the Batería III APROV correlated positively and highly (rs $\geq .75$) with the WISC arithmetic test. All four tests also correlated positively with nonverbal intelligence (rs $\geq .58$).

Test-retest reliability. As shown in Table 4, test-retest reliability was excellent for all the tests (rs $\geq .93$).

Discussion

This study validated the math tests of the Batería III APROV for use in the first six school grades in Spain. The four tests showed good psychometric properties suggesting that they are appropriate for use in Spain, at least for the ages 6 to 13 years. Except for the school Grades 4 and 5,

performance increased with age and boys performed slightly better than girls. However, school grade and sex interactions were not significant, indicating that these sex differences are independent of the maturational processes associated with age.

All tests showed a high degree of construct validity, as shown by the results of the CFAs. The bifactor EFA models that were eventually fitted for the tests Calculation and Quantitative Concepts are in line with the structure of the tests as the items are of ascending difficulty. In the bifactor models the first and easier items of each test loaded on one bifactor, whereas the last, more difficult items loaded on the second bifactor.

The ascending difficulty of the test items was also apparent in the fact that all 6-year-olds in this study answered correctly the first items of the tests Calculation, Applied Problems, and Quantitative Concepts, that are intended for the assessment of younger children.

Validity evidence was also demonstrated by the high internal consistency of the tests and by the significant effects of schooling with total scores increasing by school grade. However, no significant differences in total scores were found between the fourth and fifth grades. One interpretation of these latter findings could be that the four math tests of the Batería III APROV do not include enough items at this difficulty level to discriminate performance between the two school grades. However, these results should not necessarily be regarded as artifacts of the measurement for two reasons. First, they may indicate differences between the American and the Spanish school curriculum. For instance, in this study, no children answered Item 26 (i.e., the math problem: 3 + 6(8) = ?) in the test Calculation correctly although a significant number of children in school Grades 5 and 6 could answer correctly Items 27 to 32. Apparently, Spanish students learn certain mathematical operations later than their American peers. Second, the lack of significant differences between fourth and fifth graders could be due to developmental effects. As shown by the magnitude (i.e., effect sizes) of the differences between consecutive school grades, the findings are in line with those of previous studies showing that the most rapid academic growth in mathematics is observed in the first three school grades (Lee, 2010; Lichten, 2004; McGrew & Woodcock, 2001). In fact, additional analyses showed that although there was a main effect of school grade on WISC arithmetic performance, $F(5, 424) = 41.93$, $p < .01$, fourth and fifth graders again did not differ significantly ($d = 0.26$). Longitudinal studies that follow the same sample of students across primary and secondary education are needed to cast some light on the factors that influence the developmental trajectory of mathematical ability.

Test-retest reliability was excellent and the four tests also correlated highly with the WISC arithmetic ability and with nonverbal intelligence. It should however be noted that the WISC arithmetic ability test requires working memory and verbal abilities and, although it correlates highly, it is not a mathematic achievement test. Nevertheless, given the lack of standardized, mathematics tests in Spain, we considered the WISC arithmetic test and the K-BIT Matrices as appropriate validity criteria.

In line with previous findings (Else-Quest et al., 2010), in this study, boys, compared to girls, performed better in three out of the four mathematic achievement tests of the WJ-III ACH battery. However, the magnitude of this sex difference was small. What this study adds to the literature is that we obtained no school grade by sex effects, indicating that these gender differences are present already in the first grades of primary education. Interestingly, one study of preschool children based on the WJ-III ACH battery reported no gender differences in mathematics (Mathews, Ponitz, & Morrison, 2009), a finding confirmed also by other studies using other mathematic ability tests (e.g., Aubrey & Godfrey, 2003; Aunola, Leskinen, Lerkkanen, & Nurmi, 2004). Future studies need to examine how the transition from preschool to primary education may result in gender differences in mathematics.

Certain limitations of this study should be noted. First, we did not provide any evidence of divergent validity by including an assessment of vocabulary. Second, considering that the

participants of this study were all recruited from the area of Murcia, future studies need to verify whether results hold across other regions of Spain. Third, we did not ensure that children who were not native Spanish speakers were sufficiently proficient to participate in the study. However, all parents of nonnative Spanish children reported that they spoke Spanish at home and all children had attended a Spanish school for more than 1 year.

Conclusion. The findings of this study suggest that practitioners can feel secure in using and interpreting the mathematical achievement subtests of the Batería III APROV in the first six school grades in Spain as the tests showed good psychometric properties. However, results of students in the fourth and fifth grades should be interpreted with caution as the findings of this study indicate that the tests may not significantly discriminate ability between these two school grades. Although, in this study, several of the first items of the tests Calculation, Applied Problems, and Quantitative Concepts showed no variability in the lower grades, it is imperative that these items are included in the assessment of the mathematic abilities of children aged 6 to 8 to identify those with MLD. As there are no mathematic ability tests validated for Spain, future norming of the Batería III APROV will allow comprehensive psychological assessments of children's math ability. Future studies need also to include adolescent and adult samples to verify that the validity of the mathematics tests of the Batería III APROV holds even when all test items are included and that significant differences are found across all school grades.

Appendix

Translation of Test Words From Latin American Spanish to Spanish

Test: Applied Problems		Test: Applied Problems	
Nouns		Verbs	
crayones	colores	Te quedarían	Te quedan
centavos	céntimos	tuvieras/ tendrías	tienes
dolares	euros	consiguieras	consigues
cuadras	metros	recibes	te dan
duraznos	melocotones	señale	indica
auto	coche	compone	tiene
galón	litro		
millas	kilómetros	Test: Quantitative Concepts	
pulgadas	centímetros	Verbs	
costo	precio	debe ir	va
balance	resto	nombra	dime
término	final	muestra	significa
		encuentra	calcula

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: The first author is financed by a Juan de la Cierva research fellowship. This study is part of the COEDUCA project and was financed by Grants CSD2008-00048 and PSI2011-23340 by the Spanish Ministry of Science and Innovation.

References

Alarcón, M., Knopik, V. S., & DeFries, J. C. (2000). Covariation of mathematics achievement and general cognitive ability in twins. *Journal of School Psychology, 38*, 63-77.

Alsina, A., & Sáiz, D. (2003). Un análisis comparativo del papel del bucle fonológico versus la agenda visuo-espacial en el cálculo en niños de 7-8 años [A comparative analysis of the paper version of the phonological curl versus the visual-spatial agenda in 7-8 years-old children's calculation]. *Psicothema 15*, 241-246.

American Psychiatric Association (2000). *Diagnostic and statistical manual of mental disorders* (4th ed., text rev.). Washington, DC: Author.

Ardila, A. (1995). Directions of research in cross-cultural neuropsychology. *Journal of Clinical and Experimental Neuropsychology, 17*, 143-150.

Aubrey, C., & Godfrey, R. (2003). The development of children's early numeracy through Key Stage 1. *British Educational Research Journal, 29*, 821-840.

Aunola, K., Leskinen, E., Lerkkanen, M. K., & Nurmi, J. E. (2004). Developmental dynamics of math performance from preschool to Grade 2. *Journal of Educational Psychology, 96*, 699-713.

Barbaresi, W. J., Katusic, S. K., Colligan, R. C., Weaver, A. L., & Jacobsen, S. J. (2005). Learning disorder: Incidence in a population-based birth cohort, 1976-82, Rochester, Minn. *Ambulatory Pediatrics, 5*, 281-289.

Checa, P., Rodriguez-Bailon, R., & Rueda, M. R. (2008). Neurocognitive and temperamental systems of self-regulation and early adolescents' social and academic outcomes. *Mind Brain and Education, 2*, 177-187.

Else-Quest, N. M., Hyde, J. S., & Linn, M. C. (2010). Cross-national patterns of gender differences in mathematics: A meta-analysis. *Psychological Bulletin, 136*, 103-127.

Hu, L. T., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional versus new alternatives. *Structural Equation Modeling, 6*, 1-55.

Kaufman, A. S., & Kaufman, N. L. (1990). *Kaufman Brief Intelligence Test* (Traduccion al español, TEA Ediciones, Madrid). Bloomington, MN: Pearson.

Lee, J. (2010). Tripartite growth trajectories of reading and math achievement: Tracking national academic progress at primary, middle, and high school levels. *American Educational Research Journal, 47*, 800-832.

Lichten, W. (2004). On the law of intelligence. *Developmental Review, 24*, 252-288.

Martín, E., Martínez-Arias, R., Marchesi, A., & Pérez, E. M. (2008). Variables that predict academic achievement in the Spanish compulsory secondary educational system: A longitudinal, multi-level analysis. *Spanish Journal of Psychology, 11*, 400-413.

Mathews, J. S., Ponitz, C. C., & Morrison, F. J., (2009). Early Gender Differences in Self-Regulation and Academic Achievement. *Journal of Educational Psychology, 101*, 689-704.

McGrew, K. S. & Woodcock, R. W. (2001). *Woodcock-Johnson III Technical Manual*. Itasca, IL: Riverside Publishing.

McGrew, K. S., Schrank, F. A., & Woodcock, R. W. (2007). Technical Manual. *Woodcock-Johnson III Normative Update*. Rolling Meadows, IL: Riverside.

Miranda, A., Meliá, A., Marco, R., Roselló, B., & Mulas, F. (2006). Dificultades en el aprendizaje de matemáticas en niños con trastorno por déficit de atención e hiperactividad [Difficulties in learning mathematics in children with attention deficits and hyperactivity]. *Revista de Neurología 42*, 163-170.

Moines, L. E. S. M., Allen, D. N., Puente, A. E., & Neblina, C. (2010). Validity of the WISC-IV Spanish for a clinically referred sample of Hispanic children. *Psychological Assessment, 22*, 465-469.

Muñoz-Sandoval, A. F., Woodcock, R. W., McGrew, K. S., & Mather, N. (2005). *Bateria III Woodcock-Muñoz*. Rolling Meadows, IL: Riverside.

Muthén, L. K., & Muthén, B. O. (1998-2007). *Mplus user's guide* (5th ed.). Los Angeles, CA: Muthén & Muthén.

Organisation for Economic Co-operation and Development. (2009). *The PISA 2009 profiles by country/economy*. Retrieved from http://stats.oecd.org/PISA2009Profiles/#

Peak, L. (1996). *Pursuing excellence: A study of U.S. eighth-grade mathematics and science achievement in international context*. Washington, DC: U.S. Department of Education.

Rodriguez, V. L., & Diaz, J. O. P. (1990). Correlations among GPA and scores on the Spanish version of WISC-R and the Woodcock-Johnson achievement subtests for 10-year-old- to 12-year-old Puerto-Rican children. *Psychological Reports, 66*, 563-566.

Rosselli, M., Ardila, A., Bateman, J. R., & Guzmán, M. (2001). Neuropsychological test scores, academic performance and developmental disorders in Spanish-speaking children. *Developmental Neuropsychology, 20*, 355-373.

Shalev, R. S., Manor, O., & Gross-Tsur, V. (2005). Developmental dyscalculia: A prospective six-year follow-up. *Developmental Medicine and Child Neurology, 47*, 121-125.

Wechsler, D. (2004). *Manual for the Wechsler Intelligence Scale for Children* (4th ed., Traduccion al español, TEA Ediciones, Madrid, 2005). San Antonio, TX: Harcourt.

Wirth, R. J., & Edwards, M. C. (2007). Item factor analysis: Current approaches and future directions. *Psychological Methods, 12*, 58-79.

Woodcock, R. W., McGrew, K. S., & Mather, N. (2001, 2007). *Woodcock-Johnson III tests of achievement*. Rolling Meadows, IL: Riverside.

Woodcock, R. W., McGrew, K. S., Schrank, F. A., & Mather, N. (2001, 2007). *Woodcock-Johnson III normative update*. Rolling Meadows, IL: Riverside.

Yu, C.Y. (2002). *Evaluating cutoff criteria of model fit indices for latent variable models with binary and continuous outcomes* (Unpublished doctoral dissertation). University of California, Los Angeles.