Viapiana et al. Psicologia: Reflexão e Crítica (2016) 29:39
DOI 10.1186/s41155-016-0045-5

Psicologia: Reflexão e Crítica

RESEARCH Open Access

Development of the Arithmetic Subtest of the School Achievement Test-Second Edition

Vanisa Fante Viapiana[1*], Euclides José de Mendonça Filho[2], Rochele Paz Fonseca[1], Claudia Hofheinz Giacomoni[2] and Lilian Milnitsky Stein[1]

Abstract

The School Achievement Test (*Teste de Desempenho Escolar*, TDE) has been widely used in clinical and educational contexts for the past 22 years. Arithmetic disorders are frequent among children and teenagers, requiring new and updated tasks to assess as accurately as possible school achievement. The last decade has witnessed a growing recognition of the need for significant changes in educational assessment practices. Evidence provided by item response theory (IRT) enabled the link of more detailed information improving assessment quality. The aim of this study was to develop a revised and completely updated version of the Arithmetic Subtest for the School Achievement Test-Second Edition (*Teste de Desempenho Escolar-Segunda Edição*, TDE-II). To this end, two studies were conducted. The first study focused on item and test construction, while the second study assessed the preliminary version of the instrument. The sample consisted of 302 students in grades 1 through 9 recruited from public and private schools. Factor analysis revealed two factors which explained 74 % of the variance in the data. Both dimensions were closely related to item complexity and difficulty. The subtest was therefore divided into two versions: one for students in grades 1 through 5 and the other for those in grades 6 through 9. Both versions were analyzed based on IRT models, which suggested that the items provided a comprehensive measure of the latent trait. The results provided satisfactory evidence of internal structure and reliability. Results indicated that the Arithmetic Subtest of the TDE-II has adequate psychometric properties for the assessment of arithmetic skills in primary education. Interpretation based on IRT analyses can be helpful for future studies about math education, discriminating even better between learning difficulty and typical groups, with the data to be the basis of math cognition stimulation programs.

Keywords: School achievement, Arithmetic, Psychometric, Item response theory

Background

The School Achievement Test (*Teste de Desempenho Escolar*, TDE) is a measure of educational achievement divided into three subtests: reading, writing, and arithmetic. It was originally developed for children in grades 1 through 6 (Stein, 1994). The arithmetic subtest evaluates the skills involved in oral problem solving and the written calculation of mathematical operations (Stein, 1994).

Since its publication, the TDE has been widely used in scientific studies throughout Brazil. The Arithmetic Subtest has proved to be a useful tool for the study of relationships between mathematical learning and neuropsychological or cognitive abilities (Costa et al., 2011), the identification of children with mathematical difficulties and the selection of appropriate control groups (Oliveira-Ferreira et al., 2012), and the assessment of correlations between school performance and teacher ratings (Capellini et al. 2004). However, factors such as the recent change in the Brazilian primary school syllabus have created the need for a revision of the TDE (Knijnik et al. 2013).

Arithmetic disorders are frequent among children and teenagers, requiring new and updated tasks to assess as accurately as possible accomplishing educational changes. Although achievement tests are important tools for educational research and development, few instruments are available for the assessment of mathematical skills in the Brazilian population. Psychological and neuropsychological instruments such as the

* Correspondence: vanisaviapiana@gmail.com
[1]Pontificia Universidade Catolica do Rio Grande do Sul, Porto Alegre, Rio Grande do Sul, Brazil
Full list of author information is available at the end of the article

Wechsler Intelligence Scales (Rueda et al. 2012) and the Brief Neuropsychological Assessment Instrument (NEUPSILIN) (Fonseca et al. 2009) make use of arithmetic tasks to evaluate cognition. More specific measures of arithmetic processing such as the Arithmetic Test (Seabra et al. 2010) and the Neuropsychological Battery for the Assessment of Numerical Processing and Calculation Skills in Children-ZAREKI (Silva and Santos, 2011) are also available. However, there is a need for instruments with norms for Brazilian children and which adequately reflect the contents of the primary school curriculum (Rodrigues et al. 2010).

The development of mathematical competence is associated with several other neurocognitive mechanisms (Geary, 2004; Menon, 2010). Mathematical competence also requires extensive teaching, years of practice, and cognitive effort (Haase et al. 2015) as it evolves over the course of schooling (Seabra et al. 2010).

The school is usually responsible for the stimulation of mathematical thinking. The assessment of learning plays a fundamental role in the development of teaching and learning strategies and the implementation of more successful educational policies at school. Learning assessments are often performed through educational tests, whose results are used to monitor individual student progress, as well as evaluate schools and state education systems based on normative standards (American Education Research Association et al. 2014).

Given the great demand for educational assessment, the wide usage of the TDE, and the recent changes in the Brazilian primary school curriculum, this study aimed to develop a revised version of the Arithmetic Subtest for the second edition of the TDE. The theoretical (study 1) and empirical (study 2) procedures following those suggested by Pasquali (2010) and by the *Standards for Educational and Psychological Testing* (2014 edition) (American Education Research Association et al. 2014).

Study 1 involved the construction of items and a global test for the assessment of mathematical skills in primary school. In study 2, items were empirically tested, analyzed, and selected for the final version of the Arithmetic Subtest of the School Achievement Test-Second Edition (*Teste de Desempenho Escolar-Segunda Edição*, TDE-II).

Methods

Study 1

The Arithmetic Subtest of the TDE-II was developed in two stages. The first sought to provide a theoretical and operational definition of the construct of arithmetic, while the second involved the construction of items to assess the contents learned by students at every school year.

Stage 1

Item development was preceded by systematic textbook analysis and consultation with experts in mathematics, as suggested by guidelines on the construction of educational tests (Urbina, 2007). The contents covered during each school year were first identified based on the National Textbook Program (Programa Nacional de Livros Didáticos, PNLD) developed by the Brazilian Ministry of Education (MEC) (MEC, 2012, 2013). The PNLD provides pedagogical support to schools by evaluating the textbooks used in primary education.

Analysis of the PNLD's guidelines was performed with the help of an expert in mathematical teaching in primary school, with graduate-level training and over 20 years of classroom experience. The expert analyzed the PNLD guidelines published in 2013 and 2014, which recommend textbooks for the first and final years of primary school, respectively. By analyzing the summary of all books recommended by the Ministry of Education, the expert identified the arithmetic content covered in each book. A frequency analysis was then performed to identify the contents included in over 80 % of textbooks for each grade. The results of this frequency analyses are shown in Table 1.

These contents served as a basis for the development of items for the revised version of the TDE, described in stage 2 of this study.

Stage 2

The second stage of the study aimed to develop items to evaluate the arithmetic contents for each one of the nine elementary school years. Two teachers with graduate-level training in mathematics and with over 10 years of classroom experience in primary education helped on the development of the items. To ensure that all skill levels for each content were covered, a total of 197 items were initially developed to evaluate number recognition and writing, counting, and sequencing (covered by the mathematical literacy curriculum in grades 1, 2, and 3), as well as each of the contents taught in subsequent school years. Each content was covered by three items, ranging in difficulty from easy to hard. This strategy was established based on the potential performance of students who were halfway through the school year. Five items from the original Arithmetic Subtest of the TDE were also included, for a total of 202 items.

The suitability of each item for the assessment of arithmetic ability for Brazilian students was then examined. Items were reviewed by six mathematics teachers who served as expert judges. All participants had graduate-level training and over 5 years of professional experience. Three had experience with the first years of primary school and therefore analyzed the items developed for grades 1 through 5. The other three teachers worked with the final years of primary school and were therefore responsible for evaluating the items constructed for children in grades 6 through 9.

Table 1 Arithmetic contents for each year of the primary school curriculum

Grade	Content
1st	Counting
	Sequencing
	Addition
	Subtraction
2nd	Counting
	Sequencing
	Addition
	Subtraction
3rd	Sequencing
	Composition and decomposition
	Four natural operations
4th	Decimal numbers reported to the thousandth place
	Composition 100.000
	Four natural operations
	Fractions: notions and comparison
5th	Decimal numbers reported to the billionth place: comparison and sequencing
	Four natural operations
	Fractions: notions and comparison
	Fractions and percentages
	Fractions and decimal numbers
	Arithmetic operations with fractions and natural numbers
	Fractions: equivalence and simplification
	Numerical expressions
6th	Four natural operations
	Comparison of decimal numbers
	Fractional decimals
	Fractions: order
	Fractions: operations
	Numerical expressions
7th	Whole numbers: comparison
	Whole numbers: addition and subtraction
	Whole numbers: multiplication and division
	Ratios and proportions
	Percentages
	Numerical expressions
8th	Numerical expressions—rational numbers
	Comparison and operations involving whole, rational, irrational, and real numbers
9th	Comparison and operations involving whole, rational, irrational, and real numbers
	Exponentiation: operations
	Radicals: operations

The judges answered three questions about each of the items: two categorical questions with dichotomous answer options (yes-no): (1) Does the item cover the content area indicated? and (2) Is the item adequate for this grade?; and an additional question answered on a three-point Likert scale (easy, medium, hard): (3) How difficult is this item for the school year in question? This question intended to verify whether the judges agreed with the classifications of item difficulty established during item construction.

Inter-rater agreement was analyzed based on the frequency of each answer. Agreement mean rates ranged from 95 (items of first through fifth year) to 100 % (sixth through ninth year) on the first question, 91 (1–5 years) to 97 % (6–9 years) in the second, and 67 to 81 % in the third question (among judges responsible for grades 1 through 5 and 6 through 9, respectively). The lowest rates of agreement were observed in the assessment of item difficulty. This may be explained by a lack of standardized teaching practices, especially in the first years of primary education. We analyzed inter-rater agreement based on the content validity index (Alexander and Coluci, 2011). We chose to maintain items with 100 % inter-rater agreement on the first two questions. In relation to item difficulty, we selected items that obtained inter-rater agreement from at least two judges on the level of difficulty (66 %), since difficulty level would be also assessed empirically though the item response theory (IRT) in the second study. The maintained 133 items were evaluated by a graduate professor of mathematics education with 30 years of professional experience. This procedure aimed to assess the importance and frequency of the contents, as recommended by the *Standards for Educational and Psychological Testing* (2014 edition) (American Education Research Association et al. 2014). The expert reanalyzed the items to confirm their relevance to the contents covered between grades 1 and 9. Based on redundancy criterion, he excluded 29 items that assess the same content and level of difficulty and suggested two new items to address any gaps in skill

level. Items were placed in theoretical ascending order of difficulty. This version included Arabic number writing to the exponentiation and radicals operations, mean of nine items for each school year, yielding a preliminary version of the Arithmetic Subtest of the TDE-II containing a total of 102 items.

Study 2

Once all steps prescribed by psychometric instrument construction theory (American Education Research Association et al. 2014; Pasquali, 2010) were performed, the items were empirically tested. Study 2 aimed to analyze the construct validity of the Arithmetic Subtest and select items for the final version of the instrument.

Sample

A convenience sample of 313 students in grades 1 through 9, from the metropolitan region of Porto Alegre, Rio Grande do Sul, Brazil, completed intellectual and educational tests. Intellectual ability was used as inclusion criterion. Eleven students who obtained intellectual disability-level scores in the intelligence test (5th percentile or less on Raven Progressive Matrices) were excluded from the sample. Therefore, the final sample was composed for 302 students, 184 respondents of Raven's Colored Progressive Matrices Special Scale (Angelini et al. 1999) (mean = 26.20 points, percentile 10th to 99th) and 118 respondents of Raven's Progressive Matrices General Scale (Vilhena et al. in press) (mean = 40.68 points, percentile 6th to 98th).

Socioeconomic characteristics of the sample were analyzed according to the Brazilian Association of Research Companies (ABEP) (Associação Brasileira de Empresas de Pesquisa (ABEP) 2012) and showed an adequate distribution of the sample regarding Brazilian social class (class A1 to class E) (0.8 % A1; 8.4 % A2; 18.4 % B1; 38.1 % B2; 20.1 % C1; 10.9 % C2; 0.4 % C3; 2.5 % D). More characteristics of the 302 participants are shown in Table 2.

Table 2 Sample characteristics

		Type of school		Gender		Age	
Grade	*N*	Public	Private	Female	Male	Mean	SD
1st	24	13	11	11	13	7.0	0.43
2nd	26	14	12	13	13	7.8	0.63
3rd	32	19	13	18	14	9.0	0.96
4th	48	35	13	25	23	9.8	0.99
5th	49	22	27	21	28	10.9	0.57
6th	50	27	23	32	18	12.1	0.91
7th	31	13	18	13	18	13.0	0.63
8th	29	16	13	12	17	14.1	0.98
9th	13	0	13	5	8	14.5	0.52
Total	302	159	143	150	152	10.8	2.32

Instruments

All students were administered the preliminary version of the Arithmetic Subtest as well as an intelligence test used as exclusion criterion: Raven's Colored Progressive Matrices Special Scale (Angelini et al. 1999) or Raven's Progressive Matrices General Scale (Vilhena et al. in press). The former was administered to 184 students 11 years and 9 months old or younger, while the latter was administered to 118 students older than 11 years and 10 months.

Data collection procedures

All students received written consent for participation from parents and/or legal guardians. Intellectual performance tests were administered individually in grades 1 through 3 and collectively, in groups of eight children, from grade 4 onwards. On a second session, in the same week, students were administered the preliminary version of the Arithmetic Subtest of the TDE-II. The test was administered individually to students in grades 1 through 3 and collectively, in the classroom, in grade 4 onwards. During collective administration, all students in each classroom responded to the TDE-II Arithmetic Subtest. However, the tests completed by children who did not have parental consent for participation were later discarded.

Each student received a test containing the items developed for their own school year, as well as the two preceding years and the following grade. A fifth-grade student, for instance, responded to the items pertaining to grades 3–6. The inclusion of items pertaining to the preceding and following grades was used to confirm the difficulty of each item. The use of items from two preceding grades was also justified by the diagnostic criteria for mathematical learning disorders, which state that the child must be at least two school years behind their peers in mathematical achievement to qualify for a diagnosis of dyscalculia (Haase et al. 2011). Individual applications were interrupted after ten consecutive errors to avoid exposing children to items that they could not answer to avoid anxiety and frustration. During classroom administrations, students were asked to respond to each question as best they could and to draw a diagonal line across any items which they had not yet learned to solve.

Data collection was performed in the second semester of the school year (October and November). At this point, students were likely to have been exposed to all mathematical contents relevant to their school year.

Data analysis procedures

Data analysis was guided by the idea that arithmetic skills develop gradually over the course of schooling so that students were likely to respond correctly to items covered in the three previous grades but would not

know how to answer those covered over two grades above their own. Test items were coded as dichotomous variables (right/wrong), and those which students did not know how to answer were considered errors.

Data analysis was performed in two stages. The underlying dimensions of the subtest were first determined by principal axis exploratory factor analysis. Fit of dimensional models was evaluated considering the following: (1) *Tucker-Lewis index* (TLI) > 0.90 and (2) root mean square error of approximation (RMSEA) and *standardized root mean square residual* (SRMR) < 0.08 (Beaujean, 2014) provided by the *psych* package. At this stage, an item was considered adequate when (1) the absolute factor loading was over 0.30 for samples between 300 and 350 participants (significance at $p < 0.05$ and statistical power of 80 %); (2) if the item loaded onto two or more factors with loading differences of at least 0.10; and (3) it had an unequivocal relationship to the dimension in question, demonstrating theoretical convergence with other items in the same factor (Hair et al. 2010).

Then, item response theory (IRT) was used to estimate the latent trait (proficiency) underlying the participant responses to the test (DeMars, 2010; Embretson and Reise, 2000). This method evaluates individual differences along a developmental continuum, identifying the skill level for which the item is maximally informative, regardless of student proficiency (Laros, 2005; Hambleton and Swaminathan, 2010). IRT also enables to asses specific discrepancies between observations and model predictions, known as residuals, to evaluate whether the intended inferences made from the model are trustworthy (Glas, 2016).

Likelihood ratio tests between fitted nested IRT models were performed comparing the Rasch model (all item discriminations constrained to one), the one-parameter model (1PL; item discriminations are the same but freely estimated), and the two-parameter model (2PL). Significant likelihood ratio tests, log-likelihood (the bigger the better), Akaike's information criterion (AIC; the smaller the better), and the Bayesian information criterion (BIC; the smaller the better) indicate better fit of nested models (Glas, 2016; Rizopoulos, 2006), which endorsed model selection. Then, residual analysis of the chosen model was investigated. Item-fit analysis was used to determine whether the model fits for the individual items. In this case, the statistic is calculated by comparing the frequency of individuals correctly responding to the item for a given (very small) range of values of θ (proficiency), with the model-predicted frequency (Finch and French, 2015). The resulting statistic is distributed as a chi-square. However, under many conditions, it does not in fact conform to the chi-square distribution, making it improper for testing model fit. Therefore, an alternative approach using Monte Carlo procedure was employed to obtain robust item-fit tests (Finch and French, 2015). Items with significance at the 0.05 level were considered inappropriate. M2 statistics for dichotomous data fit (Maydeu-Olivares and Joe, 2006) were considered for overall model fit; at this point, a model is considered adequate when the comparative fit index (CFI) and TLI > 0.90 and RMSEA and SRMSR < 0.8. Finally, the remaining items with difficult and discrimination standard errors bigger than one were deleted, according with cutoff criteria suggested by Hambleton and Swaminathan (2010).

It is important to notice that, unfortunately, there is no unique rule of thumb for estimating sample size requirements for IRT (DeMars, 2010). Empirical studies have suggested that sample sizes of 200 people can provide robust parameter estimates for the 2PL model (Primi and Nunes, 2005), and Monte Carlo simulation studies have shown that as few as 250 examinees satisfy IRT requirements (Embretson and Reise, 2000). Thus, we considered our sample size ($N = 302$) adequate for the estimation of difficulty and discrimination of TDE-II items.

Data were analyzed using the R Software (R Core Team, 2015), and functions implemented by the packages *psych* (Revelle, 2015) for exploratory factor analysis and *mirt* (Chalmers, 2012), *ltm* (Rizopoulos, 2006), and mokken (van der Andries, 2007) for item response theory analysis were used.

Results

Preliminary descriptive analyses led to the exclusion of four items (A45, A94, A101, A102) for lack of variability. The remaining 98 items had a KMO of 0.94 and Bartlett's test of sphericity was also significant, χ^2 (4753) = 21185.82, $p < 0.0001$, confirming that the items were appropriate for factor analysis. Dimensionality was assessed using exploratory factor analysis of items' tetrachoric correlations with varimax rotation.

One-dimensional versus two-dimensional models were investigated. Analysis suggested better fit of the two-dimensional model (fit indexes: TLI = 0.99, RMSEA = 0.04, and SMRS = 0.04) against the one-dimensional model (fit indexes: TLI = 0.88, RMSEA = 0.28, and SMRS = 0.20). Both dimensions were closely related to item complexity and difficulty and explain 74 % of the data variance. Loadings on the first factor ranged from 0.55 to 0.90 ($M = 0.73$, SD = 0.08) Additional file 1, while loadings onto the second factor ranged from 0.56 to 0.91 ($M = 0.75$, SD = 0.10). Ten items were excluded due to cross-loading or low factor loadings, as previously described.

The first factor consisted of items of Arabic number writing, counting, sequencing, simple operations, and fractions. The second factor was composed of multidigit operations; fractions; decimal numbers; percentages; exponentiations; radicals; operations and comparison of the whole, rational, irrational, and real numbers; and numerical expressions.

Based on these results, the TDE-II Arithmetic Subtest was divided into versions A (first through fifth grade) and B (sixth to ninth grade). As recommended by DeMars (2010) and Laros (2005), all subsequent analyses were performed independently for each dimension. The responses of all participants were considered in both sets of items; thus, we had the response patterns of students of the sixth to ninth grades at the items of the first through fifth grades and vice versa. These procedures also identified a need for more difficult items for fifth-grade students (version A). Although more difficult items were loaded on the version B factor, some of these items were also included in version A. This criterion was adopted to ensure the test was suitable for students with high arithmetic ability. Similarly, some easier items of the fourth and fifth years were also included in version B to allow the assessment of students with poor arithmetic skills. Albeit these items correspond to different factors, the intent of these changes was to have a range of item difficulty levels required by IRT analysis. It is important to note that the items from different factors were adequate for both versions of the test.

Psychometric properties of version A (first through fifth grade) of the Arithmetic Subtest of the TDE-II

Version A was initially composed of 56 items. A principal axis factor analysis of the tetrachoric correlation matrix identified a single factor which explained 67 % of the variance in scores (KMO = 0.95, Bartlett's test of sphericity χ^2 (1326) = 12729.6, $p < 0.0001$; Cronbach's alpha = 0.95). Factor loadings ranged from 0.42 to 0.97, while the items from version B included in this measure ranged from 0.72 to 0.83.

Mokken analysis was used to evaluate test scalability (van der Linden and Hambleton, 1997). In this procedure, values of at least 0.30 justify the implementation of a parametric IRT model. Items in version A had a diagnostic value of 0.86 and, as such, were entirely suitable for parametric analysis.

Three parametric IRT models were used in data analysis (Rasch, 1PL, and 2PL). Likelihood ratio tests between fitted nested IRT models showed that the 2PL model (log-likelihood = −3387.07, AIC = 6994.14, and BIC = 7402.29) had better fit since it showed lower residuals than the Rasch (log-likelihood = −4210.63, AIC = 8531.26, and BIC = 8735.34) and 1PL models (log-likelihood = −3483.34, AIC = 7078.68, and BIC = 7286.4). Subsequently, item fit of the 2PL model was analyzed, in which four items were deleted due to misfit (A25, A29, A48, A62). Finally, 11 items (A1, A5, A6, A8, A12, A60, A61, A64, A65, A66, and A67) presented standard errors of difficulty or discrimination bigger than one and were excluded from the model.

The identification of the most discriminating items for each level of difficulty (Embretson and Reise, 2000) led to the selection of items which were more discriminating but still allowed for the assessment of a continuum of arithmetic ability. Thus, in order to reduce the length of version A, six items (A18, A21, A29, A41, A37, and A39) were excluded. Additional content analysis was then performed to verify whether the 35 remaining items covered the latent trait investigated. After initial assessment, the items were presented to a mathematics expert, who identified a lack of intermediate-difficulty items to assess multiplication and subtraction. Two items were therefore included in the subtest, and their association with the remaining items was investigated, resulting in adequate item fit. The resulting version of the instrument contained 37 items. This version consisted of number processing; addition and subtraction in oral problems (three items); Arabic number writing (two items); counting (three items); sequencing (two items); simple operations: addition (five items), subtraction (six items), multiplication (five items), and division (four items); number composition (one item); and fractions: notions/comparisons/operations (six items). The 37 items had a log-likelihood of −2567.54, AIC of 5279.08, and a BIC of 5546.23. The new set of items was not statistically significant at the item-fit test, thus fitting adequately the data, and the overall M2 fit indexes of the model were as follows: CFI = 0.98, TLI = 0.98, RMSEA = 0.04, and SRMSR = 0.05. Item difficulty ranged from −3.31 to 0.94, as shown in Table 3.

The IRT scores of first- through fifth-grade students on this measure ranged from −2.45 to 0.97, with a mean of 0.56 (SD = 0.81). For this skill range, version A provides 85.4 % of the total information provided by the 2PL model, as shown by the test information curve (Fig. 1). This suggests that, for the sample of first- to fifth-grade students, all responses feel into the discriminative range of the test; therefore, the model is capable of differentiating the sample of students along the ability continuum. The results indicated the increase of the score means in each subsequent school year (first M = −1.71, SD = 0.38; second M = −1.26, SD = 0.37; third M = −0.84, SD = 0.59; fourth M = −0.12, SD = 0.49; fifth M = 0.18, SD = 0.36). One-way ANOVA showed a statistically significant difference between groups (F(4, 174) = 101.75, $p < 0.00$). Tukey post hoc test confirmed a statistically significant difference between the means of the first and second grades (mean difference = −0.45, $p < 0.01$, Cohen's d = 1,24), second and third grades (mean difference = −0.41, $p < 0.01$, Cohen's d = 0,83), third and fourth grades (mean difference = −0.72, $p < 0.001$, Cohen's d = 1.36), and fourth and fifth grades (mean difference = −0.32, $p <$ 0.01, Cohen's d = 0.73).

Table 3 Difficulty and discrimination power of items in version A of the TDE-II

Items	Difficulty	Discrimination	Items	Difficulty	Discrimination
A2	−3.3375	1.5039	A28	−0.6929	3.4331
A7	−3.2336	1.6934	A36	−0.6369	3.6589
A18	−2.7796	2.4006	A33	−0.4758	5.27
A13	−2.7351	1.9641	A27	−0.4697	4.9345
A9	−2.3072	3.113	A37	−0.4043	3.9365
A3	−2.1235	2.0514	A39	−0.3681	4.4655
A15	−2.0947	2.5557	A31	−0.3645	8.5817
A10	−1.9997	4.0321	A53	0.1308	3.4349
A16	−1.8	2.2701	A44	0.2164	3.5976
A17	−1.7922	3.2982	A54	0.3288	3.3836
A14	−1.6235	2.7877	A58	0.5078	4.0837
A19	−1.5748	3.1029	A50	0.5188	3.151
A20	−1.4458	5.147	A59	0.5835	4.5851
A22	−1.415	2.5523	A52	0.6275	3.8564
A23	−1.1939	2.9703	A56	0.7888	3.2923
A26	−1.0552	5.0518	A63	0.8019	4.596
A24	−0.9618	4.2261	A57	0.8577	4.4113
A30	−0.8141	3.8369	A55	0.922	2.2936
A32	−0.7083	3.2816			

Psychometric properties of version B (grades 6 through 9) of the Arithmetic Subtest of the TDE-II

Version B was initially composed of 56 items and, like version A, contained a single underlying factor explaining 68 % of the variance in scores (KMO = 0.93, Bartlett's test of sphericity χ^2 (1540) = 16298.96, $p < 0.0001$; Cronbach's alpha = 0.97). Factor loadings ranged from 0.51 to 0.97, with items drawn from version A showing loadings of 0.73 to 0.95. Item scalability was 0.78, which allowed for the use of parametric IRT models.

In version B, the 2PL model (log-likelihood = −3714.92, AIC = 7653.84, and BIC = 8069.41) also produced a better fit to the data than the Rasch (log-likelihood = −4333.53, AIC = 8779.05, and BIC = 8986.83) and 1PL models (log-likelihood = −3837.66, AIC = 7789.33, BIC = 8000.82).

Goodness of fit to the 2PL model was analyzed following the same procedures described for version A. Three items (A93, A66, and A95) presented standard errors of difficulty or discrimination bigger than one and were excluded. The remaining 53 items were not statistically significant at the item-fit test, thus fitting adequately the data. Then, ten items (A36, A48, A58, A78, A55, A76, A80, A83, A75, and A91) were excluded due to content redundancy. The final set of 43 items was again analyzed and approved by the mathematics expert. Version B was composed of multidigit operations: addition (1 item), subtraction (2 items), multiplication (3 items), and division (2 items); fraction notions/comparisons/orders/operations (13 items); decimal numbers and fraction transformation (4 items); decimal number comparison (2 items); whole number operations (3 items); percentages (2 items); exponentiation (2 items); radicals (1 item); numerical expressions (6 items); and comparison involving whole, rational, irrational, and real numbers (2 items). Its final version ranged in difficulty from −0.64 to 3.21, as shown in Table 4. The scores obtained by the sixth- through ninth-grade students ranged from −0.94 to 2.68 ($M = 0.87$,

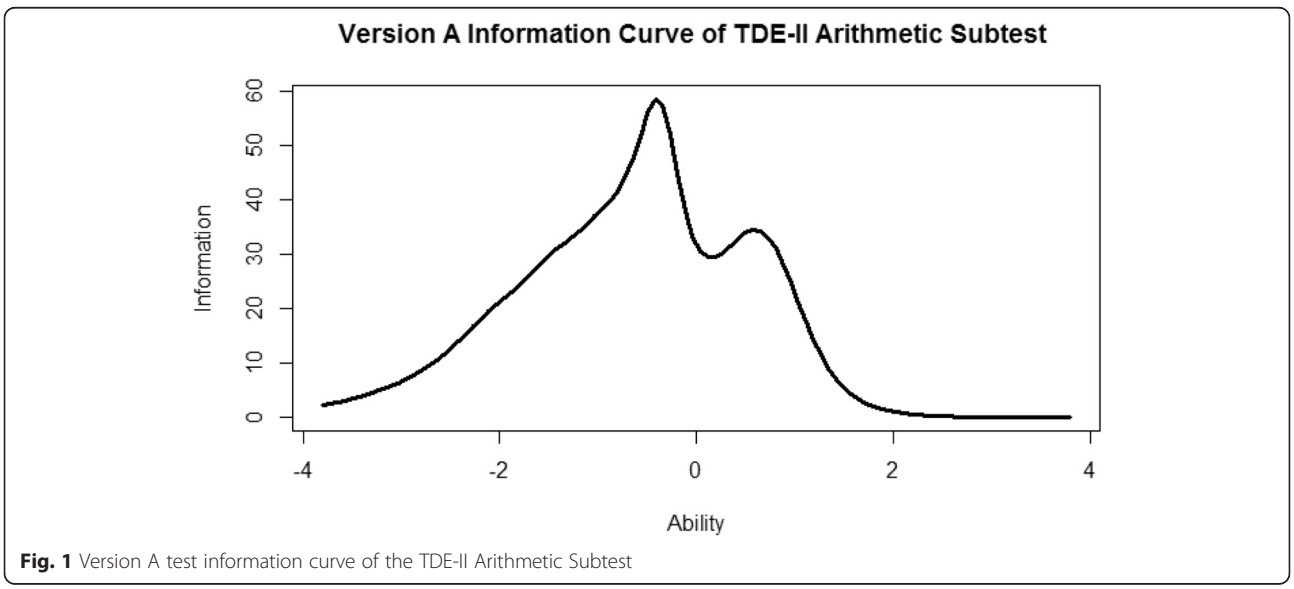

Fig. 1 Version A test information curve of the TDE-II Arithmetic Subtest

Table 4 Difficulty and discrimination power of items in version B of the TDE-II

Item	Difficulty	Discrimination	Item	Difficulty	Discrimination
A38	−0.6484	3.0287	A67	1.1098	5.716
A37	−0.6369	3.3532	A77	1.1859	1.4303
A41	−0.6065	3.4624	A88	1.2549	2.0311
A39	−0.604	4.1499	A90	1.2574	3.2208
A53	−0.0067	3.2261	A82	1.2785	3.8193
A62	0.064	2.5025	A72	1.3361	2.6694
A44	0.0858	3.2814	A69	1.4662	4.0853
A54	0.2088	3.3916	A84	1.6088	2.7979
A50	0.4287	2.7401	A89	1.7578	2.5416
A59	0.5039	4.1512	A86	1.7958	2.4449
A46	0.5444	3.6426	A96	1.8705	3.2276
A52	0.5649	3.046	A85	1.8709	2.3781
A56	0.7565	2.6032	A70	1.946	2.4555
A63	0.7584	4.0741	A68	2.0229	2.2619
A57	0.8137	4.1131	A71	2.045	3.2069
A87	0.8208	2.6471	A97	2.1589	3.8918
A65	0.9048	6.3845	A73	2.1863	2.2984
A64	0.9239	6.278	A98	2.5193	2.446
A61	0.9441	5.2496	A81	2.7061	2.6243
A60	0.9588	5.3662	A100	2.7857	1.7694
A79	1.0391	3.0221	A92	3.2164	1.8715
A74	1.0473	2.846			

SD = 0.76) in the final 2PL model (log-likelihood = −2904.77, AIC = 5981.54, and BIC = 6300.64). All items were not significant at the residual analysis, and the M2 fit indexes were as follows: CFI = 0.98, TLI = 0.98, RMSEA = 0.05, and SRMSR = 0.07.

Version B provided 91.01 % of information for this skill range, as shown by the test information curve (Fig. 2). This finding suggests that for sixth- to ninth-grade students, all responses fitted into the discriminative range of the test. The results indicated the increase of the score means in each subsequent school year (sixth $M = 0.28$, SD = 0.54; seventh $M = 0.73$, SD = 0.44; eighth $M = 1.57$, SD 0.33; ninth $M = 1.84$, SD = 0.76). One-way ANOVA test showed a statically significant difference between the groups ($F(3, 119) = 69.11$, $p < 0.001$). Tukey post hoc test confirmed a statistically significant difference between the means of the sixth and seventh (mean difference = −0.45, $p < 0.01$, Cohen's $d = 0.82$) grades and seventh and eighth grades (mean difference = −0.83, $p < 0.01$, Cohen's $d = 0.91$). However, the difference between the eighth and ninth grades was not statistically significant (mean difference = −0.27, $p = 0.08$).

Discussion

The systematic procedures involved in the construction of the Arithmetic Subtest of the TDE-II were performed in two studies. The results revealed that the TDE-II Arithmetic Subtest has adequate psychometric properties for the assessment of mathematical ability in primary school students.

The first study described the theoretical definition of the construct based on textbooks recommended by the Brazilian Ministry of Education. The association between the main contents of primary arithmetic studied in Brazilian schools and the content covered by the test provide important evidence of content and construct validity (American Education Research Association et al. 2014; Pasquali 2009). Additionally, the analysis by mathematics experts ensured that the intended construct was adequately captured throughout the test development process (Pasquali, 2009). The use of national mathematics textbooks as guidelines suggests that the test may have appropriate content validity to evaluate arithmetic skills throughout the entire country.

The selection of two-parameter models in study 2 is in accordance with the literature on the analysis of instruments containing items of widely varying difficulty (DeMars, 2010, Laros, 2005). Although the TDE-II is based on educational principles, the analysis of the cognitive demand associated with different mathematical operations reveals several mechanisms underlying task performance. The demand on cognitive resources such as the executive functions, which include working memory, varies as a function of the type and complexity of the mathematical task, as well as the age of the child/adolescent (Geary, 2006; Purpura and Ganley 2014; Bull and Lee, 2014). Our results showed that the subtest could be divided into simple and complex tasks, as suggested by Menon's arithmetic information processing model (2010). According to Menon's model, complex mathematical computations differ from simple ones in that they place greater demands on executive functions, while simpler ones depend mostly on the recall of arithmetic information from memory.

The underlying test structure revealed by factor analysis suggested that items increase significantly in complexity over the course of up to over the fifth grade. These findings confirm the idea that fifth grade marks the consolidation of the arithmetic concepts taught in early primary school for the subsequent development of more complex abilities (Rodrigues et al. 2010).

According to the textbooks recommended by the Brazilian Ministry of Education, the teaching of fractions begins in the fourth grade—when the basic idea of fractions is first introduced—and continues in the fifth grade, with the study of arithmetic operations with fractions. However, these concepts are often covered in

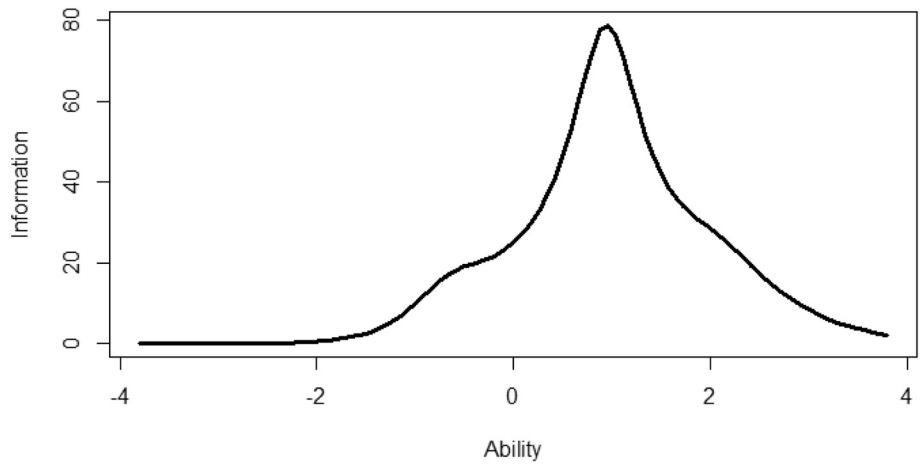

Fig. 2 Version B test information curve of the TDE-II Arithmetic Subtest

the final chapters of different textbooks and, as reported by the expert and some participating judges, are often relegated to later grades.

According to the IRT, both versions of the subtest had an adequate range of difficulty and were sensitive to individual differences in the sample. The information provided by each set of items is illustrated by the test curves in Figs. 1 and 2, allowing for an assessment of the ability of each version to reflect the continuum of arithmetic ability displayed in the sample (Hambleton and Swaminathan, 2010). As shown in the two figures, versions A and B provide a large amount of information on the abilities which they intend to measure (arithmetic skills in the first- through fifth-grade students and sixth- through ninth-grade ones, respectively). Lastly, scores appeared to increase over time, suggesting that arithmetic development follows a relatively linear course. This provides evidence of the discriminating potential of each item for every school year.

Some limitations in study 2 should be considered, such as the sample size and the fact that all participants were recruited from the metropolitan region of Porto Alegre, Rio Grande do Sul. Given the preliminary nature of the present study, we suggest that future studies look further into the psychometric properties of this instrument. To ensure that a test is adequately interpreted, there is a need for multiple sources of validity evidence (American Education Research Association et al. 2014), in addition to reliability, standardization, and normative comparison procedures.

Conclusions

The Arithmetic Subtest of the TDE-II has adequate psychometric properties for the assessment of mathematical ability over the course of primary school. It therefore addresses the gap in the Brazilian literature regarding updated instruments to evaluate mathematical ability in school-age children. The developed tool based on IRT analyses can be helpful for future studies about math education, discriminating even better clinical groups with learning difficulty and typical groups, with data to be the basis of math cognition stimulation programs.

Additional file

Additional file 1: Table S1. Factor loadings of the Arithmetic Subtest TDE-II regarding two dimensions. (DOCX 24 kb)

Abbreviations

IRT, item response theory; PNLD, Programa Nacional de Livros Didáticos; TDE-II, *Teste de Desempenho Escolar-Segunda Edição*

Authors' contributions

VFV: She worked in the idealization of the study, literature review, data collection, result analysis, and discussion and in the general wording of the article. EJMF: He was responsible for the data analysis, reported the results of the second study, and contributed in the discussion of these results. RPF: She contributed to the overall design of the study and reviewed all the procedures of the test development, as well as the intellectual content of the article. CHG: She contributed to the idealization and overall design of the study, directed all the psychometric procedures, and revised the intellectual content of the article. LMS: She contributed to the idealization and overall design of the study and reviewed all the procedures of the test development, as well as the intellectual content of the article. All authors read and approved the final manuscript.

Competing interests

The authors declare that they have no competing interests.

Author details

[1]Pontificia Universidade Catolica do Rio Grande do Sul, Porto Alegre, Rio Grande do Sul, Brazil. [2]Universidade Federal do Rio Grande do Sul (UFRGS), Porto Alegre, Brazil.

Received: 22 July 2016 Accepted: 4 August 2016
Published online: 18 August 2016

References

Alexander, N. M. C., & Coluci, M. Z. O. (2011). Validade de conteúdo nos processos de construção e adaptação de instrumentos de medidas. *Ciência & Saúde Coletiva*, 16(7).doi: 10.1590/S1413-81232011000800006.

American Education Research Association, American Psychological Association, & Nacional Council on Measurement in Education. (2014). *Standards for educational and psychological testing*. Washigton: AERA, APA, NCME.

Angelini, A. L., Alves, I. C. B., Custódio, E. M., Duarte, W. F., & Duarte, J. L. M. (1999). *Manual matrizes progressivas coloridas de Raven: escala especial*. São Paulo: Centro Editor de Testes e Pesquisas em Psicologia.

Associação Brasileira de Empresas de Pesquisa (ABEP). (2012). *Critério de Classificação Econômica Brasil*. Brasil. Retrieved from: file:///C:/Users/User/Downloads/09_cceb_2014%20(3).pdf

Beaujean, A. A. (2014). *Latent variable modeling using R: a step-by-step guide*. New York: Taylor & Francis. Routledge. doi: 10.4324/9781315869780

Bull, R., & Lee, K. (2014). Executive functioning and mathematics achievement. *Child Development Perspectives, 8*(1), 34–41. doi:10.1111/cdep.12059.

Capellini, S. A., Tonelotto, J. M. F. & Ciasca, S. M. (2004) Medidas de desempenho escolar: avaliação formal e opinião de professores. *Estudos de Psicologia,* 21 (2), 79-70. Retrieved from: http://www.scielo.br/pdf/estpsi/v21n2/a06v21n2.pdf.

Chalmers, R. P. (2012). mirt: a multidimensional item response theory package for the R environment. *Journal of Statistical Software, 48*(6), 1–29. Retrieved from http://www.jstatsoft.org/v48/i06/paper\npapers3://publication/uuid/0258AE87-10FF-45E9-BF61-631B467ECFB4.

Costa, A. J., Lopes-Silva, J. G., Pinheiro-Chagas, P., Krinzinger, H., Lonnemann, J., Willmes, K., Wood, G., & Haase, V. G. (2011). A hand full of numbers: a role for offloading in arithmetics learning. *Frontiers in Psychology, 2*, 1–12. doi:10.3389/fpsyg.2011.00368.

DeMars, C. (2010). *Item response theory*. Lodon: Oxford University Press.

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. Mahwah, NJ: Lawrence Erlbaum Associates.

Finch, W. H., & French, B. F. (2015). *Latent variable modeling with R*. New York, NY: Routledge.

Fonseca, R. P., Salles, J. F., & Parente, M. A. M. P. (2009). *Instrumento de Avaliação Neuropsicológica Breve—NEUPSILIN* (1ªth ed.). São Paulo: Vetor Editora.

Geary, D. C. (2004). Mathematics and learning disabilities. *Journal of Learning Disabilities, 37*, 4–15. doi:10.1177/00222194040370010201.

Geary, D. C. (2006). Development of mathematical understanding. In W. Damon (Ed.) & D. Kuhl & R. S. Siegler (Vol. Eds.) Handbook of child psychology (6th ed): *Cognition, perception, and language.* Vol 2 (pp. 777-80). New York: John Wiley & Sons.

Glas, C. A. W. (2016). Frequentist model-fit tests. In W. J. van der Linden (Ed.), *Handbook of item response theory, volume two: statistical tools* (pp. 313–361). Boca Raton, FL: CRC Press.

Haase, V. G., Moura, R. J., Pinheiro-Chagas, P. & Wood, G. (2011). Discalculia e dislexia: semelhança epidemiológica e diversidade de mecanismos neurocognitivos. In L. M. Alves, R. Mousinho& S. A. Capellini, S. A. (Eds.) *Dislexia: novos temas, novas perspectivas* (pp. 257-282). Rio de Janeiro: Wak.

Haase, V. G., Ferreira, F. O., de Moura, R. J., Pinheiro-Chagas, P., & Wood, G. (2015). Cognitive neuroscience and math education: teaching what kids don't learn by themselves. *Jornal Internacional de Estudos em Educação Matemática, 5(2).*

Hair, J. F., Black, W. C., Babin, B. J., & Anderson, R. E. (2010). Multivariate data analysis. New York, NY: Pearson Prentice Hall.

Hambleton, R. K., & Swaminathan, H. (2010). *Item response theory: principles and applications.* Norwell, MA: Kluwer Nijhoff Publishing

Knijnik, L. F., Giacomoni, C. H., & Stein, L. M. (2013). Teste de Desempenho Escolar: um estudo de levantamento. *Psico-USF, 18*(3), 407–416. doi:10.1590/S1413-82712013000300007.

Laros, J. A. (2005). O uso da análise fatorial: algumas diretrizes para pesquisadores. In L. Pasquali (Ed.), *Análise fatorial para pesquisadores* (pp. 163–184). Brasilia: LabPAM.

Maydeu-Olivares, A., & Joe, H. (2006). Limited information goodness-of-fit testing in multidimensional contingency tables. *Psychometrika, 71*(4), 713–732. doi:10.1007/s11336-005-1295-9.

Menon, V. (2010). Developmental cognitive neuroscience of arithmetic: implications for learning and education. *Mathematics Education, 42*, 515–525. doi:10.1007/s11858-010-0242-0.

Ministério da Educação. (2012) *Guia de Livros Didáticos PNLD 2013: Alfabetização Matemática e Matemática—Ensino Fundamental anos Iniciais*. Brasília: Secretaria de Educação Básica. Retrieved from: http://www.fnde.gov.br/programas/livro-didatico/guias-do-pnld/item/3773-guia-pnld-2013-%E2%80%93-ensino-fundamental.

Ministério da Educação. (2013) *Guia de Livros Didáticos PNLD 2013: Matemática—Ensino Fundamental Anos Finais*. Brasília: Secretaria de Educação Básica. Retrieved from: http://www.fnde.gov.br/programas/livro-didatico/guias-do-pnld/item/4661-guia-pnld-2014.

Oliveira-Ferreira, F., Costa, D. S., Micheli, L. R., Oliveira, L. F. S., Pinheiro-Chagas, P., & Haase, V. G. (2012). School achievement test: normative data for a representative sample of elementary school children. *Psychology & Neuroscience, 5*(2), 157–164. doi:10.3922/j.psns.2012.2.05.

Pasquali, L. (2009). Psicometria. *Revista da Escola de Enfermagem da USP*, *43*, 992-999. doi: 10.1590/S0080-62342009000500002.

Pasquali, L. (2010). *Instrumentação psicológica: fundamentos e práticas*. Porto Algre: Artmed. 568 p.

Primi, R., & Nunes, C. H. S. S. (2005). Impacto Do Tamanho Da Amostra Na Calibração De Itens E Estimativa De Escores Por Teoria De Resposta Ao Item. *Avaliação Psicológica*, *4*(2).

Purpura, D. J., & Ganley, C. M. (2014). Working memory and language: skill-specific or domain-general relations to mathematics. *Journal of Experimental Child Psychology, 122*, 104–121. doi:10.1016/j.jecp.2013.12.0090022.

R Core Team. (2015). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. http://www.R-project.org/.

Revelle, W. (2015). *psych: procedures for personality and psychological research*. Evanston, Illinois. Retrieved from http://cran.r-project.org/package=psych Version = 1.5.8.

Rizopoulos, D. (2006). ltm: an R package for latent variable modelling and item response theory analyses. *Journal of Statistical Software, 17*(5), 1–25. Retrieved from: http://www.jstatsoft.org/v17/i05/.

Rodrigues, S. D., Guassi, A. R., & Ciasca, S. M. (2010). Avaliação do desempenho em matemática de crianças do 5° Ano do ensino fundamental estudo preliminar por meio do teste de habilidade matemática (THM). *Revista Psicopedagogia, 27*(83), 181–190. Retrieved from: http://pepsic.bvsalud.org/scielo.php?pid=S0103-84862010000200004&script=sci_arttext.

Rueda, F. J. M., Noronha, A. P. P., Sisto, F. F., Santos, A. A. A., & Castro, N. R. (2012). *Escala de Inteligência Wechsler para Crianças—WISC-IV*. São Paulo: Casa do Psicólogo.

Seabra, A. G., Dias, N. D., & Macedo, E. C. (2010). Desenvolvimento das Habilidades Aritméticas e Composição Fatorial da Prova de Aritmética em Estudantes do Ensino Fundamental. *Revista Interamericana de Psicologia., 44*, 481–488. Retrieved from: http://www.redalyc.org/articulo.oa?id=28420658010.

Silva, P. A., & Santos, F. H. (2011). Discalculia do Desenvolvimento: Avaliação da Representação Numérica pela ZAREKI-R. *Psicologia: Teoria e Prática, 27*(2), 169–177. doi:10.1590/S0102-37722011000200003.

Stein, L. M. (1994). *TDE—Teste de Desempenho Escolar: manual para aplicação e interpretação*. São Paulo, SP: Casa do Psicólogo.

van der Andries, A. L. (2007). Mokken scale analysis in R. *Journal of Statistical Computing*, *20*(11).

van der Linden, W. J., & Hambleton, R. K. (1997). *Handbook of modern item response theory*. New York, NY: Springer.

Vilhena C., Guntert. I. B., & Tosi, A. D., (in press). *Teste Matrizes Progressivas de Raven*. São Paulo: Casa do Psicólogo.

Urbina, S. (2007). *Fundamentos da testagem psicológica*. Porto Alegre: Artmed.