European Journal of Science and Mathematics Education, 2022, **10**(3), 352-365
ISSN: 2301-251X (Online)

OPEN ACCESS

Research Article

Preliminary Development of an Item Bank and an Adaptive Test in Mathematical Knowledge for University Students

Fernanda Belén Ghio [1*]
0000-0002-4223-2470

Manuel Bruzzone [1]
0000-0002-7782-7802

Luis Rojas-Torres [2]
0000-0002-9085-2703

Marcos Cupani [1]
0000-0003-2132-5552

[1] Instituto de investigaciones Psicológicas (IIPSI), Consejo Nacional de Investigaciones Científicas y Técnicas (CONICET), Facultad de Psicología, Universidad Nacional de Córdoba (UNC), ARGENTINA
[2] Institute for Psychological Research, University of Costa Rica, COSTA RICA
[*] Corresponding author: fernandabghio@gmail.com

Citation: Ghio, F. B., Bruzzone, M., Rojas-Torres, L., & Cupani, M. (2022). Preliminary Development of an Item Bank and an Adaptive Test in Mathematical Knowledge for University Students. *European Journal of Science and Mathematics Education, 10*(3), 352-365. https://doi.org/10.30935/scimath/11968

ARTICLE INFO

Received: 16 Aug 2021
Accepted: 24 Mar 2022

ABSTRACT

In the last decades, the development of computerized adaptive testing (CAT) has allowed more precise measurements with a smaller number of items. In this study, we develop an item bank (IB) to generate the adaptive algorithm and simulate the functioning of CAT to assess the domains of mathematical knowledge in Argentinian university students (N=773). Data were analyzed from the Rasch model. A simulation design created with the R software was used to determine the necessary items of the IB to estimate examinee ability. Our results indicate that the IB in the domains of mathematical knowledge is adequate to be applied in CAT, especially to estimate average ability levels. The use of CAT is recommended for rapidly generating indicators of the knowledge acquired by students and to design educational strategies that enhance student performance. Results, constrains, and implications of this study are discussed.

Keywords: Rasch model, item bank, domains of mathematical knowledge, adaptive test, higher education, simulation

INTRODUCTION

Mathematical literacy forms an integral part of the basic content in the plan of studies of diverse disciplines, such as physical, medical, and social sciences (Lindquist et al., 2017), and in different educational levels, namely, primary, secondary, and university education (Doran, 2017; Scheffield, 2005). At university, gaining mathematical competencies and skills usually generates difficulties, manifested in student low achievement in mathematics (Rodriguez et al., 2015). Precise assessment measurements are required to identify mathematical content either acquired or not by students and to generate strategies to improve the teaching-learning process and the educational quality of institutions (Flores & Gómez, 2009).

As mentioned, poorly developed mathematical performances have been reported in all educational levels. In relation to this, Engelbrecht et al. (2007) identified that, in technical secondary schools, students showed deficits in learning mathematical procedures, as a result of their unfamiliarity with content area questions in the educational course. Similarly, mathematics tests were found to cause greater anxiety compared with

other subject areas (e.g., language, science) (Putwain et al., 2010), and in specific university studies such as engineering (Karjanto & Yong, 2013).

A system that has become relevant in the last decades is the one consisting of large-scale assessments with the structure of computerized adaptive testing (CAT) (Chang, 2015). These standardized tools combine technology and innovation in education (Lee Bouygues, 2019). In contrast to traditional assessment tests, with CAT, each examinee is presented with a tailor-made test which determines the level of the student knowledge in a specific subject area, reducing test times/length, maximizing measurement precision and efficiency, delivering immediate results, and helping to significantly decrease costs (Barrada et al., 2006).

In addition, CAT is particularly useful in distance education and evaluation (Kaya & Tan, 2014). This assessment tool is designed to adapt their level of difficulty in view of the responses provided by test-takers, thus matching the examinee's knowledge and ability. If a student gets their question wrong, the test will follow up with an easier question, but if the student gets their question right, the following question will be more challenging (Costa & Ferrao, 2015); thus, ensuring equitable outcomes (Aybek & Demirtasli, 2017; Han, 2018). Diverse studies indicate that CAT reduces by up to 50% the number of items and administration time, providing measurements with a higher level of precision with a lower number of items (Kingsbury & Houser, 1999; Wainer, 2000).

In particular, studies have identified that assessment methods affect students' performance in mathematics (Pollock, 2002), and there is evidence that students who respond to CAT, on average, perform better than those who take traditional assessments (Čisar et al., 2016). In addition, this type of test is useful for answering content questions, e.g., mathematics. It also helps reduce the level of test anxiety in mathematics as the system adapts to the student's ability level (Linacre, 2000). CAT is developed from psychometric models (usually one- or two-parameter models) based on the item response theory (IRT). Using the Rasch model allows determining which items in a test can be solved by an examinee (Baker & Kim, 2004). This enables designing intervention strategies that help examinees without ability or without the required knowledge of the content assessed in the items. Furthermore, it allows placing items and persons along the same latent continuum (unit of measurement logit) with the purpose of developing and calibrating a set of items with varying degrees of difficulty to assess different ability levels (Čisar et al., 2010).

The development and calibration of an Item Bank (IB) are the first steps to design CAT. These procedures imply that the set of items must be applied to a sample of participants to establish its psychometric properties from the Rasch model or from other IRT model. This set of items must have adequate psychometric properties; otherwise, the results obtained would not be reliable enough (Tseng, 2016). Sometimes, the items from IBs are administered in sets distributed in different forms; in these cases, the difficulty parameters of all items must be placed in a common metric (a procedure called equating). The mean-sigma method (Navas, 1996) is one of the most commonly used equating procedures due to its simplicity and stable estimation parameters. This method defines the values of the constants of the slope of line (k) and intercept (d), using the mean (M) and standard deviations (SD) of the difficulty parameters of the anchor items (Kolen & Brennan, 2014).

The performance of CAT is established from predetermined rules. First, we define test starting rules, e.g., delimiting difficulty level at the beginning of the test. Then, we determine how the next items will be selected. Finally, we specify estimation method and criterion for ending item administration (Phankokkruad, 2012). One of the strategies applied when designing CAT involves the use of simulation studies in the different stages of its development (planning, building, quality control of the items, and algorithm development; Han, 2018). This procedure allows a faster and more economic approximation to datasets than that provided by the actual application of CAT. It also allows researchers to fix the examinee true ability level and collect data with a lower level of bias and standard errors (Olea et al., 1999); thus, obtaining quality information about the IB developed.

Although several studies conducted in Argentina report the use IRT models for the development and calibration of IBs assessing different domains of knowledge (e.g., biology, law, history, and literature) in university students (Cupani et al., 2016, 2017; Ghio et al., 2019), there is not yet an IB which evaluates the domains of mathematical knowledge. Hence, in this work we propose the development of an IB to be applied in adaptive tests in domains of mathematical knowledge, with content covering the first years of the programs of studies required in higher education. This research aims at:

(a) formulating a set of items in the domains of mathematical knowledge;

(b) applying and calibrating the items in these domains of knowledge;

(c) developing the algorithm of the adaptive test; and

(d) assessing the number of items needed/required to estimate ability levels with the IB created, using a simulation design.

MATERIALS AND METHODS

Participants

The sample consisted of 773 participants, 38.6% (298) female, 56.8% (439) male, and 4.6% (36) unreported sex, with an age range of 18-58 (M=20.14, SD=4.83). Participants were students from the School of Economics (FCE-UNC) (26%), School of Exact, Physical and Natural Sciences (FCEFyN-UNC) (47.2%), School of Mathematics, Astronomy and Physics (FAMAF-UNC) (1.9%), from National University of Córdoba (UNC), National Technological University (UTN) in Córdoba (24.1%), and Aeronautical University Institute (IUA) (0.8%). At the time of the test, students were attending subjects from the first (31.4%), second (28.3%), third (17.5%), fourth (3.1%), and fifth (0.6%) year of their degree program of studies, respectively; 19% of the students did not report the year of attendance. Participants responded to 271 questionnaires (Form A), 251 questionnaires (Form B), and 251 questionnaires (Form C) comprising the test.

Procedure

Prior consent was given by the different academic departments. Tests were administered collectively during regular class time under the supervision of the teacher in charge. Students were informed that questionnaires would be completed anonymously and voluntarily; that the test consisted of 30 multiple-choice questions with one correct option only; and that it would take 50 to 60 minutes/and that it would take no longer than 60 minutes.

Instrument

Item bank of the general knowledge test–Domains of mathematical knowledge

Different activities to ensure the adequate building of an instrument must be performed so that the test measure precisely and reliably the domains it is supposed to measure (Downing & Haladyna, 2006; Rojas-Torres & Ordóñez, 2019). To design the item pool and response model, we carried out the following tasks:

a. **Content analysis and specification table:** We analyzed mathematics content in the syllabi of 11 curricula from different university schools and colleges, in the city of Córdoba, Argentina. With the collected information, we created a spreadsheet (Excel) specifying program, year of attendance to which the program belongs to, general content, and specific content. Then, we performed a frequency analysis and consulted experts (professors of this programs) in mathematical contents about the pertinence of the selected contents and their difficulty level, which resulted in the contents with the highest importance: functions, derivatives, integration, integers, complex numbers, and real numbers, among others (**Table 1**). The experts identified frequent or infrequent content in the curriculum at that level, and main or secondary content that should be acquired by the students. Informative items were drawn up on the basis of mathematics content and degree of difficulty (Cupani et al., 2016).

Table 1. Content and number of items included according to the level of importance of the specific content area for the domains of mathematical knowledge

Content	Number of items	Percentage (%)	Content	Number of items	Percentage (%)
Function	16	14.81	Function variation	2	1.85
Matrix	11	10.19	Mathematical logics	2	1.85
Integral	9	8.33	Complex number	1	0.93
Vector space	9	8.33	Continuity	1	0.93
Derivative	8	7.41	Circumference	1	0.93
Equation	7	6.48	Set	1	0.93
Series	7	6.48	Counting	1	0.93
Vector	5	4.63	Graph	1	0.93
Real number	5	4.63	Field	1	0.93
Limit	4	3.70	Operation	1	0.93
Straight line &plane	3	2.78	Mathematical induction	1	0.93
Determinant	3	2.78	Addition	1	0.93
Integer	3	2.78	Curve	1	0.93
Succession & series	2	1.85	Topology	1	0.93

Note. Total number of items=108; Total percentages (%)=100

b. **Writing and development of items:** Items were developed by content specialists, namely, mathematics professors. They received training in how to design multiple-choice questions (Moreno et al., 2004). Questions were arranged in cards including: identification code, concept, assessed cognitive category, correct choice, bibliographical source and name of the person who wrote the question. In addition, item writers were asked to provide a subjective assessment of the level of difficulty (easy, medium, and difficult) of each item according to the level of knowledge and ability required, thus allowing item calibration during the phase of test design, assembly, and production. This commonly reduces examinee exposure to difficult items at the beginning of the test. In total, 108 items were developed, each with three options and one correct answer. Following numerous studies, we devised three-option multiple-choice items, which reduce test length without affecting test measurement accuracy and psychometric quality. Moreover, the reduce test length allows include new content in the assessment. The writers don´t always manage to write more than three plausible, but incorrect, options and also do not guarantee controlling the guessing effect (Gierl et al., 2017; Haladyna & Rodriguez, 2013; Rodriguez, 2005). All test items were thoroughly reviewed, revised, and edited to conform to test structure and format, meet pre-determined test criteria and ensure grammaticality, readability, and information validity and reliability. People trained in item writing and editorial control reviewed the items written to analyze aspects of content, style, format, writing, and grammar according to the guidelines for writing multiple-choice items.

c. **Test design, assembly, and production:** 60 items were selected out of the 108 initial items. The selection was performed considering the most pertinent level 1 content and the rating provided by the expert judges (**Table 2**). Three versions of the test were made: Form A (30 items), Form B (30 items), and Form C (30 items). The classification of items into forms was made according to both: increasing degree of difficulty and content area, with answer options randomly presented. Anchor items (15 items) and free items (15 items) were established in each form. The three versions had the same format: a) a booklet with questions in double-sheet format, and b) a response protocol to organize the examinee scores with spaces for the response choice (A, B, or C).

Table 2. Number of items per form (A, B, and C) and per specific content area of mathematics domains

Content	Number of items (%)			Content	Number of items (%)		
	Form A (%)	Form B (%)	Form C (%)		Form A (%)	Form B (%)	Form C (%)
Function	6 (20.00)	7 (23.33)	4 (13.33)	Limit	-	-	1 (3.33)
Matrix	4 (13.33)	5 (16.67)	6 (20.00)	Determinant	1 (3.33)	-	-
Integral	3 (10.00)	5 (16.67)	5 (16.67)	Integer	-	-	1 (3.33)
Vectorial space	3 (10.00)	1 (3.33)	3 (10.00)	Succession & series	1 (3.33)	-	1 (3.33)
Derivative	2 (6.67)	3 (10.00)	-	Function variation	-	1 (3.33)	1 (3.33)
Equation	2 (6.67)	1 (3.33)	3 (10.00)	Complex number	1 (3.33)	-	-
Series	2 (6.67)	1 (3.33)	2 (6.67)	Counting	1 (3.33)	1 (3.33)	1 (3.33)
Vector	2 (6.67)	1 (3.33)	-	Mathematical induction	-	-	1 (3.33)
Real number	2 (6.67)	3 (10.00)	1 (3.33)	Addition	-	1 (3.33)	-

Note. Total number of items per form is 30 and percentage (%) is 100

Data Analysis

Rasch model: Analysis of the items in the mathematical domains

The software RUMM2030 (Andrich et al., 2010) was used for data analysis of the Rasch model. We inspected and reported the fulfilment of uni-dimensionality assumptions, local independence, person and item fit, reliability, and differential item functioning (DIF) for each version of the test.

Uni-dimensionality: We followed the method proposed by Smith (2002) to determine instrument uni-dimensionality using principal component analysis (PCA) for residuals. Observing the first residual factor, we delimited two item groups, one with positive charges >.30 and another with negative charges >.30. Paired t-tests were used to determine the presence of significant differences between the estimation of persons in both item groups. This assumption requires that the percentage of tests outside the range from -1.96 to 1.96 does not exceed 5% (Smith, 2002).

Local Independence: The dependence between items was assessed through the correlation matrix of the residues. Those correlations ≥.2 indicate associations between the items (Andrich et al., 2010).

Global fit of data to the Rasch model, items, and persons: Data behave as expected for the model when the M is close to 0 and the SD is close to 1. Besides, there is an adequate fit when, observing the behavior of items and persons, the standardized residual statistics remains within the values +/-2.5 (Pallant & Tennant, 2007). The Chi-square (χ^2) statistics/test was also used to indicate that the data fitted to the model. We obtained an χ^2 value for the total items of each form and, in turn, we used the χ^2 of each item to determine the individual fit of the item to the model. A significant χ^2 Bonferroni fit <.05 indicates mismatch of data to the Rasch model, compromising invariance (Tennant & Conaghan, 2007). Data fit to the model when the χ^2 value is low and the p-value >.05 (Cavanagh & Waugh, 2011).

Reliability index: The Person's separation index (PSI) is the measure of the internal consistency of the instrument, where .70 is an optimal value in group assessment, and .85 in individual assessment (Tennant & Conaghan, 2007).

Differential item functioning (DIF): DIF was analyzed in relation to participant sex. Differential functioning occurs when different subgroups (e.g., women and men) with the same ability level display a different behavior in their responses (Tennant & Conaghan, 2007). The characteristic curves of items were analyzed to identify DIF. An item was defined to present DIF when the test of hypothesis associated to the DIF index was significant, at a level of 5%.

Equating of the items forming the IB of the mathematics test: The mean-sigma method was used to place, in a common metric, the difficulty parameters of the set of items of the three versions of the mathematics domain test with adequate psychometric properties. Equating is always performed from anchor items, equating the scores to the version with the best psychometric properties.

Development and simulation of the algorithm for the adaptive test: To determine the number of items in the mathematics test needed to estimate the examinee ability, we used a design of simulations generated with the software R (R Core Team, 2017). This design consisted of estimating 1,000 times, for each

interest ability level, a process of simulation of responses to the algorithm of the adaptive test implemented in the mathematics test IB.

The simulation of responses to the algorithm of the adaptive test for an interest ability level j consisted in the following process:

(a) selection of the initial item;

(b) random generation of response 1, using a Bernoulli distribution with 1=1±1.702;

(c) selection of a new item considering response 1;

(d) random generation of response 2, applying the process of response 1; and

(e) repeating steps (c) and (d) until fulfilling the end condition of the adaptive algorithm.

After the 1,000-simulation process estimation for an interest ability level, we obtained:

(a) a vector with 1,000 final ability estimations through the adaptive test algorithm and

(b) a vector with 1,000 values indicating the number of items used for the estimations of those abilities.

The ability levels considered were: -1.5, -1.25, -1, -0.75, -0.5, -0.25, 0, 0.25, 0.5, 0.75, 1, 1.25, and 1.5. The end condition of the algorithm was the scope of a standard error of .5; if the condition was not fulfilled with 35 items, the algorithm estimated ability with the responses gathered. The initial item was the closest to the ability level 0.

RESULTS

Item Analysis of the Mathematics Domains from the Rasch Model

Analyses of each model (A, B, and C) were performed from the Rasch model. **Table 3** shows a summary of the results of the initial study and the final study. The items of the final study formed the IB of the test of the mathematics domains (level I).

Table 3. Summary of the initial and final versions of the fit of the mathematical knowledge test

	Residual fit of items		Residual fit of persons		Item-trait interaction		PSI	Uni-dimensionality	Items	n
	M	SD	M	SD	X^2 (df)	p		Per C < 5%		
Form A										
Initial	0.21	1.37	0.02	0.95	173.74 (90)	0.00	.60	0.00	30	271
Final	0.18	0.97	-0.00	0.90	139.11(81)	0.00	.65	0.00	27	269
Form B										
Initial	0.20	1.24	-0.04	0.96	208.04(90)	0.00	.55	0.00	30	251
Final	0.17	1.18	-0.03	0.93	199.11(87)	0.00	.57	0.00	29	249
Form C										
Initial	0.28	1.33	-0.00	0.90	178.68(90)	0.00	.57	3.31	30	251
Final	0.24	1.17	-0.01	0.89	157.95(87)	0.00	.59	2.11	28	250

Note. M: Media; SD: Standard deviation; X^2: Chi-square; df: Degrees of freedom; Per C<5%=±1.96; n: Participants per form

Form A

Items 25 (anchor), 6 (free), and 26 (free) were eliminated from the initial analysis because they did not fit the model; the participants with a response pattern contrary to those of the examinees with a similar ability level (cases 86 and 152) were also eliminated. After these modifications, the results of the final analysis indicate that the uni-dimensionality assumption (Per C <5%) was confirmed. A correlation >0.2 between item 5 (integrals) and item 2 (matrices; 0.208) was observed in the correlation matrix of the residues; however, we decided to keep that item in the IB. The M and SD values for items and persons were close to 0 and 1, showing that the empiric data fitted the Rasch model. The global chi-square test was significant [$\chi2(81)=139.11$, p=0.01], indicating that it did not fit the Rasch model. In the item-by-item analysis, only item 5 did not fit the chi-square test [$\chi2(3)=25.19$, p<.001] with Bonferroni fit <.05. The PSI was .65, revealing mean reliability. No DIF was observed in relation to participant sex (**Table 3**). **Figure 1** shows the distribution of item difficulty and people's ability in Form A.

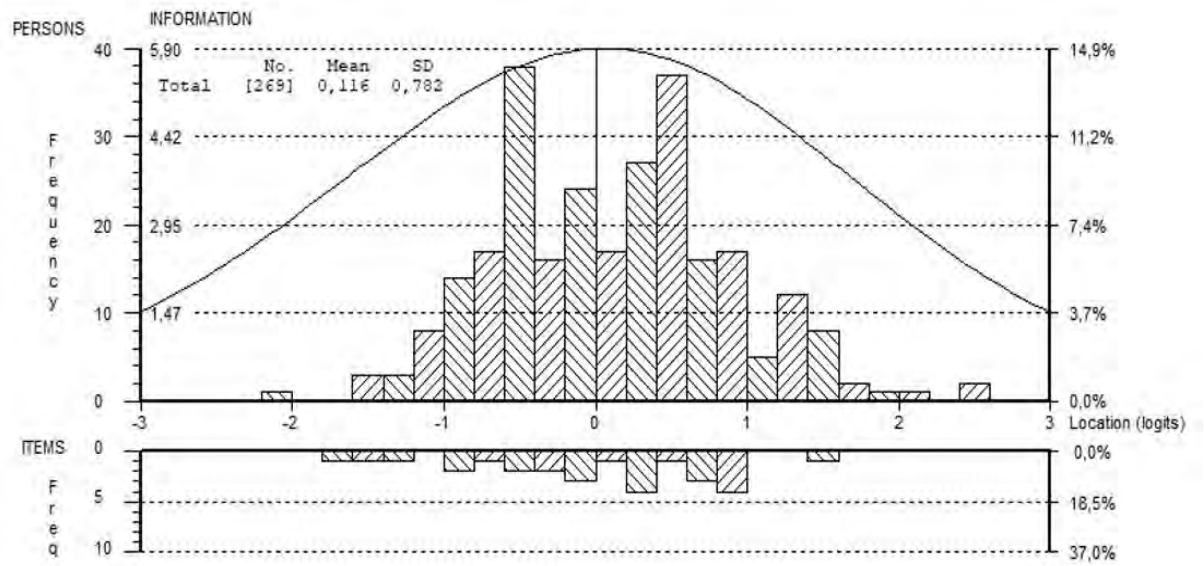

Figure 1. Distribution of item difficulty and people's ability in Form A

Form B

After the initial analysis, we decided to eliminate item 25 (anchor) and two participants who did not fit the model. The uni-dimensionality assumption and local independence were met. In the final analysis, the M and SD of items and persons agreed with the expected values. A significant chi-square [$\chi 2(87)=199.11$, $p<.01$] was observed. The fit indices for each item showed that all items fitted the residual value ±2.5. However, in the chi-square test, items 24, 28, and 29 (Bonferroni <.05) did not fit it. A low reliability index was obtained (.57). No DIF was observed in participant sex (**Table 3**). **Figure 2** shows the distribution of item difficulty and people's ability in Form B.

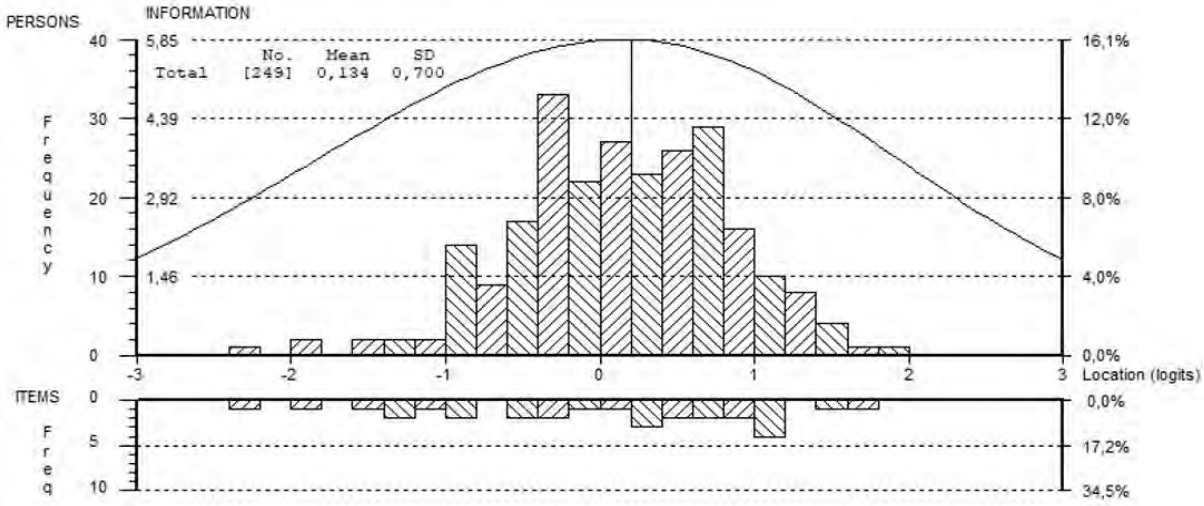

Figure 2. Distribution of item difficulty and people's ability in Form B

Form C

Items 7 and 16 and one from participant responses were eliminated. In the final analysis, the uni-dimensionality analysis was conducted. In the correlation matrix of the residuals, a correlation >.2 between item 20 (functions) and item 12 (variation of functions) of (.27) was observed. However, we decided to keep that item. The M and DF for items and persons were close to 0 and 1. The total chi-square test was significant [$\chi 2(84)=157.95$, $p<.01$]; because of that, we could not establish invariance through the trait. When considering the fit index, item 30 presented a residual value >2.5 (2.54). There was no fit in the chi-square index. The PSI

was .59. There was no DIF (**Table 3**). **Figure 3** shows the distribution of item difficulty and people's ability in Form C. From the 60 items used, 55 were kept for the test of mathematics domain IB.

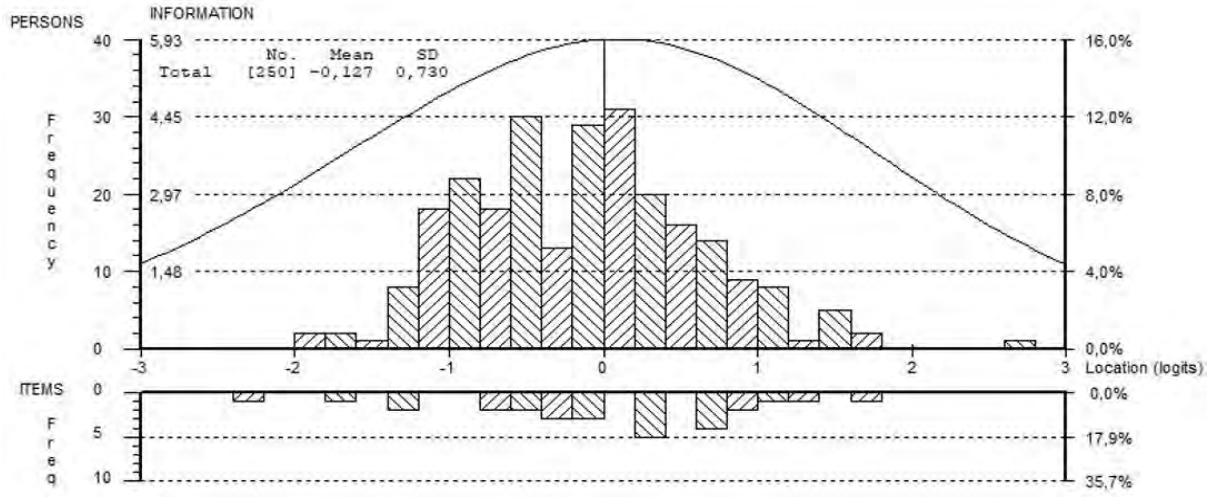

Figure 3. Distribution of item difficulty and people's ability in Form C

Equating of the Items Forming the IB of the Mathematics Test

The IB consisted of 55 items. The items were taken from the three versions considered in the study. The difficulties of Form A and Form C were equated to those of Form B, which was the one with the highest number of items with good psychometric properties. Fourteen anchor items were used; an anchor item (25) was eliminated as it did not present adequate psychometric properties (**Table 4**).

Table 4. Item difficulty equated to Form B of the mathematics domain test

Form B			Form A				Form C			
Item	*b*	Content	Item	*b* original	*b* equated	Content	Item	*b* original	*b* equated	Content
1	*-1.904*	Functions	*1*	*-1.644*	*-1.904*	Functions	*1*	*-2.39*	*-1.904*	Functions
2	*-1.397*	Matrices	*2*	*-1.279*	*-1.397*	Matrices	*2*	*-1.656*	*-1.397*	Matrices
3	*0.689*	Functions	*3*	*0.839*	*0.689*	Functions	*3*	*0.707*	*0.689*	Functions
4	*0.139*	Matrices	*4*	*0.21*	*0.139*	Matrices	*4*	*-0.267*	*0.139*	Matrices
5	*-0.817*	Integrals	*5*	*-0.954*	*-0.817*	Integrals	*5*	*-1.264*	*-0.817*	Integrals
6	-0.417	Functions				Functions	6	0.62	0.857	Vectorial spaces
7	-1.577	Integrals	7	0.742	0.940	Vectorial spaces				Equations
8	0.856	Series	8	-0.071	-0.007	Derivatives	8	1.719	1.865	Real numbers
9	0.857	Vectors	9	0.361	0.496	Series	9	-0.312	0.002	Matrices
10	-0.58	Real numbers	10	-1.407	-1.564	Real numbers	10	-0.152	0.148	Series
11	*-0.138*	Functions	*11*	*-0.093*	*-0.138*	Functions	*11*	*-0.739*	*-0.138*	Functions
12	*0.515*	Functions	*12*	*0.269*	*0.515*	Functions	*12*	*0.295*	*0.515*	Functions
13	*-0.889*	Matrices	*13*	*-0.835*	*-0.889*	Matrices	*13*	*-1.305*	*-0.889*	Matrices
14	*1.162*	Integrals	*14*	*0.69*	*1.162*	Integrals	*14*	*0.847*	*1.162*	Integrals
15	*1.031*	Counting	*15*	*0.829*	*1.031*	Counting	*15*	*0.761*	*1.031*	Counting
16	-0.347	Integrals	16	-0.577	-0.597	Determinants				Series
17	-2.233	Derivatives	17	0.039	0.121	Successions & series	17	-0.12	0.178	Limits
18	-1.265	Addition	18	-0.751	-0.800	Vectors	18	1.39	1.563	Mathematical induction
19	0.39	Real numbers	19	0.773	0.976	Complex numbers	19	-0.781	-0.429	Successions & series
20	0.34	Function variations	20	0.587	0.759	Series	20	0.935	1.146	Function variations
21	*1.148*	Real numbers	*21*	*0.828*	*1.148*	Real numbers	*21*	*0.786*	*1.148*	Real numbers

Note. *b*: Difficulty; Anchor items in *italics*

Table 4 (continued). Item difficulty equated to Form B of the mathematics domain test

Form B			Form A				Form C			
Item	*b*	Content	Item	*b* original	*b* equated	Content	Item	*b* original	*b* equated	Content
22	*-0.279*	Equations	*22*	*-0.184*	*-0.279*	Equations	*22*	*-0.504*	*-0.279*	Equations
23	*1.056*	Integrals	*23*	*0.961*	*1.056*	Integrals	*23*	*1.061*	*1.056*	Integrals
24	*0.704*	Matrices	*24*	*0.332*	*0.704*	Matrices	*24*	*0.386*	*0.704*	Matrices
26	0.296	Vectorial spaces				Vectorial spaces	26	-0.232	0.075	Vectorial spaces
27	0.485	Derivatives	27	-0.251	-0.217	Vectors	27	-0.554	-0.220	Integrals
28	1.474	Matrices	28	1.428	1.739	Equations	28	-0.021	0.269	Integrals
29	1.711	Functions	29	-0.344	-0.325	Derivatives	29	0.238	0.506	Matrices
30	-1.012	Functions	30	-0.499	-0.506	Functions	30	0.338	0.598	Vectorial spaces

Note. *b*: Difficulty; Anchor items in *italics*

Development and Simulation of the Algorithm for the Adaptive Test

Table 5 shows the results found in this analysis. The ability levels between -.75 and .50 displayed the highest levels of precision. Mean ability estimation traits were ≤.15, and SD was <.65. Moreover, most of the simulations could estimate examinee ability with less than 35 items. The value θ=.50 was the one showing the lowest percentage of cases that could not reach the established standard error with less than 35 items, with a percentage of barely 1.8%. The mean of items used to estimate examinee ability within the range of ability between -.75 a .50 was from 18 to 19 items, with SD from 1.5 to 3 items, approximately. This result indicates that with 22 items, it is possible to obtain estimations of examinee ability, with true abilities between -.75 and .50 and a standard error ≤.50 units.

Table 5. Descriptive statistics of the results of an adaptive test with the parameters of the mathematics domain test in a simulation of 1,000 examinees according to ability level

TA	MEA	DEA	MT	MnI	DnI	NS-35I
-1.50	-1.15	1.16	-0.35	25.86	7.44	342
-1.25	-0.87	1.12	-0.38	24.41	7.07	251
-1.00	-0.56	0.98	-0.44	21.02	4.09	14
-0.75	-0.60	0.58	-0.15	19.09	2.88	16
-0.50	-0.39	0.45	-0.11	18.11	1.47	2
-0.25	-0.21	0.44	-0.04	18.00	1.47	1
0.00	-0.07	0.62	0.07	18.49	2.89	5
0.25	0.18	0.46	0.07	17.95	1.46	2
0.50	0.35	0.64	0.15	18.63	3.13	19
0.75	0.55	0.87	0.20	20.61	4.72	58
1.00	0.71	0.75	0.29	20.43	4.80	65
1.25	0.95	0.93	0.30	22.74	6.39	169
1.50	1.21	0.76	0.29	23.26	6.02	151

Note. TA: True ability; MEA: Mean of estimated ability; DEA: Deviation of estimated ability; MT: Mean of trait; MnI: Mean of used items; DnI: Deviation of the mean of items; NS-35I: Number of simulations that finished with 35 items

In ability levels higher than the previous central range (from -.75 to .50), the mean values of the trait ability estimate were between .20 and .30; on the other hand, in the lowest ability levels, trait values were between .35 and .44. Besides, in the range of higher abilities, there was more precision than in the lowest range. In the first, the SD of the estimated abilities ranged from .75 to .93, whereas in the second, they were between .98 and 1.16. In the cases of -1.00, .75, and 1.00, most simulations used less than 35 items to estimate the ability level (93.5% of the simulations showed the lowest case: θ=1.00). Here, the mean number of items used ranged from 20 to 21, with an SD between 4 and 5, indicating that with approximately 26 items, it is possible to achieve a true estimation of the examinee ability. Yet, it presents high trait values and variability of ability estimates, especially in θ=-1.00.

On the other hand, in the cases of abilities ≥1.25, we obtained approximately 15% of simulations that did not reach the expected standard error with 35 items. In these cases, the mean number of used items was approximately 23, with an SD of approximately 6 items. When ability levels ranged from –1.50 to -1.25, the percentages of use of the 35 items were very high: 25.1% and 34.2%, respectively. Here, the mean numbers of used items were 24.41 and 25.86, with SDs of 7.07 and 7.44, respectively. However, the number of

simulations requiring the use of 35 items shows that this model of adaptive test is not recommended to estimate exact ability values in examinees with true abilities <-1.25.

Finally, **Figure 4** shows the mean of items used in the estimation of examinee ability according to true ability. The lowest number of items used can be found in the central values, with an increase toward the tails and a steepening in the left tail.

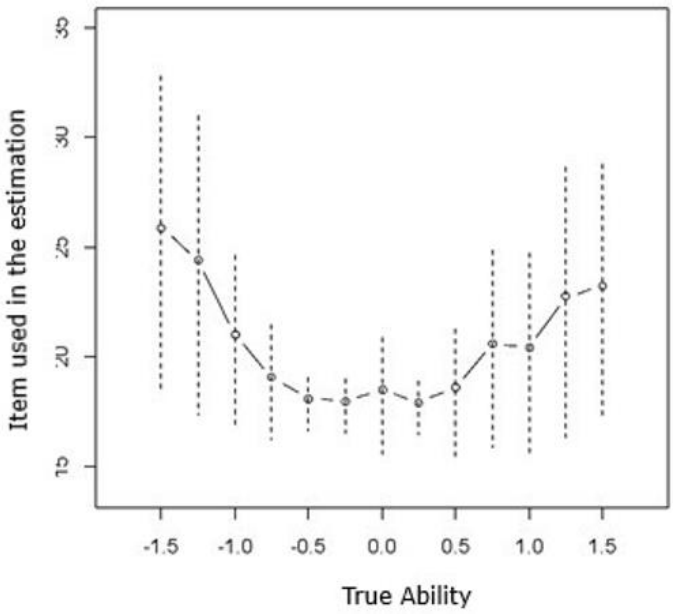

Figure 4. Mean of items used in the estimation of examinee ability according to the level of true ability. Dashed lines indicate the standard deviation of the items used

DISCUSSION

In this study, we proposed the preliminary development of an IB in the domains of mathematical knowledge, with appropriate content for university education, to be applied in CAT. The IB was analyzed and calibrated from the Rasch model to determine the fit of data to the model and the item parameters assessing mathematical content (Čisar et al., 2010). In general, the items showed adequate psychometric properties. Those items that did not fit the initial analysis (four free items and one anchor item) were eliminated; hence, the IB consisted of 55 items in mathematical knowledge (level I). In the simulation of the CAT algorithm, this IB was more precise in the medium ability levels. This implies that, to estimate ability levels below or above ± 1.00 logit, 22 to 25 items will be needed and to estimate the ability level of students with logit between -0.75 to + 0.75, the CAT algorithm will require 18 to 20 items. The analysis of the uni-dimensionality assumption provides insight into the validity of the test construct. Results show that the uni-dimensionality assumption was confirmed in the three versions of the mathematics test (Form A, Form B, and Form C), namely, the items measured a single latent construct. Such assumption is essential for developing the IB because, for scores to be comparable, equating (through anchor items) must measure a single construct (Dorans & Kingston, 1985).

The local independence assumption was confirmed in Form B, but there was a residual correlation >.2 between two items in Form A and two items in Form B, showing a possible local dependence in these pairs of items (Reeve et al., 2007). Andrich and Marais (2019) indicate which causes of such dependence must be analyzed in context; in other words, the features of the items and the purpose of the test must be considered, that is why, in our case, we could hypothesize that the correlation between the pairs of items could be established because they have a similar statement format. It could also exist because, although they assess the same type of content, the items respond to different difficulty levels (Chen & Thissen, 1997; Yen, 1993), or because some sort of halo effect might be taking place (Andrich & Marais, 2019). Anyway, we decided to keep these items to avoid under representativeness in the difficulty levels assessed.

When considering reliability, the values shown in the present study were low. Similar findings were reported in a study on the development of a diagnostic test in mathematics in Costa Rica. In that research,

reliability indices improved in subsequent applications of the test when the IB was enhanced with items covering a higher difficulty range (Zamora Araya, 2015). Our results could show that the test includes either too easy or too difficult items, or that there is a lack of items covering certain ability spectra, namely, few items provide information of certain ability levels.

The simulations performed in the present work show that the IB allows an ability estimation with an acceptable precision (standard error <.5) for examinees whose true ability is in the central values of the scale (from -1 to 1), using less than 26 items in most cases. This number of items is lower than that applied in the forms used to develop the IB in mathematical content (30); in some cases, it required a number of items as low as 14.

For persons with levels outside the central range of abilities, estimations are not that precise. However, we can conclude, with a few items, that the examinee ability is >1.25 or <-1.25. These conclusions can be commonly drawn with paper-and-pencil tests. Nevertheless, to take advantage of the benefits associated to the use of CAT, the IB must include some items with difficulty levels >1.25 and some others with difficulty levels <-1.25. This procedure will allow estimating with high precision examinee abilities with true abilities outside the central range.

Finally, to obtain estimations more accurate than those with a standard error .5, the IB must contain a high number of items representing each ability interval, derived from a very fine partition of the relevant ability continuum (from -4 to 4). This decision will depend on the uses given to the IB; in diagnostic cases, an acceptable precision may be sufficient, but in uses linked to high consequences (selection of students for educational programs), high precisions are required (Messick, 1989).

Drawbacks and Constraints

The results of the present study exhibit some constraints. On the one hand, the sample consisted of students of varied academic programs of study; however, most studied at the FCEFyN-UNC. Accordingly, in future studies, the number of students from that school should be enlarged to allow estimating high and low ability levels in the domains of mathematical knowledge. On the other hand, the items contained in the IB were not extremely precise to certain ability levels. It is necessary to enhance the IB with items covering such ability spectra, mainly items with high and low difficulty levels. Such strategy would allow enhancing the range of students' ability measured by this test and, consequently, improving the IB reliability indices. Likewise, further research should be carried out on the local dependence found in two pairs of items and define their continuity in the present IB.

The development of CAT for the domains of mathematical knowledge would allow reducing considerably the administration time of a test, facilitating its application for large samples of students, mainly in large-scale studies (Baldasaro et al., 2013). In addition, statistics indicate that students postpone the completion of their university studies and only a low percentage graduates. This is particularly common in degree programs that require mathematical knowledge (e.g., engineering, mathematics, applied mathematics, physics, computer science) (Universidad Nacional de Córdoba. Secretaría de Asuntos Académicos. Programa de Estadística Universitaria, 2020). Therefore, this is a useful tool for identifying student difficulties when undertaking university studies. This allows thinking strategies to reinforce mathematical knowledge in secondary education (Programa Estado de la Nación, 2011). Moreover, it would allow achieving immediate results of student knowledge and identifying easily problem content areas to design suitable educational strategies (Vie et al., 2017). Also, it would serve as an admission process for educational programs that require mathematical skills. In this form, students with skills in line with the curriculum for which they enrolled would be selected (Rojas et al., 2018).

CONCLUSIONS

The IB consisting of 55 items in the domains of mathematical knowledge for university students provides adequate psychometric properties to be used in CAT. The simulation process taking place in this study allowed identifying the flaws found in a real application of CAT. The development of this type of instrument will provide educational systems with new ways of incorporating technology in assessment processes.

Author contributions: All authors were involved in concept, design, collection of data, interpretation, writing, and critically revising the article. All authors approve final version of the article.

Funding: The authors received no financial support for the research and/or authorship of this article.

Declaration of interest: Authors declare no competing interest.

Data availability: Data generated or analyzed during this study are available from the authors on request.

REFERENCES

Andrich, D., & Marais, I. (2019). *A course in Rasch measurement theory: Measuring in the educational, social and health sciences*. Springer. https://doi.org/10.1007/978-981-13-7496-8

Andrich, D., Sheridan, B., & Luo, G. (2010). Rasch models for measurement: RUMM2030 [computer software]. *RUMM Laboratory Pty Ltd*. https://www.rasch.org/rmt/rmt114d.htm

Aybek, E. C., & Demirtasli, R. N. (2017). Computerized adaptive test (CAT) applications and item response theory models for polytomous items. *International Journal of Research in Education and Science, 3*(2), 475-487. https://doi.org/10.21890/ijres.327907

Baker, F. B., & Kim, S. H. (2004). *Item response theory: Parameter estimation techniques*. Marcel Dekker.

Baldasaro, R. E., Shanahan, M. J., & Bauer, D. J. (2013). Psychometric properties of the mini-IPIP in a large, nationally representative sample of young adults. *Journal of Personality Assessment, 95*(1), 74-84. https://doi.org/10.1080/00223891.2012.700466

Barrada, J. R., Olea, J., Ponsoda, V., & Abad, F. J. (2006). Estrategias de selección de ítems en un test adaptativo informatizado para la evaluación de Inglés escrito. [Item selection rules in a computerized adaptive test for the assessment of written English]. *Psicothema* [*Psychothema*]*, 18*(4), 828-834.

Cavanagh, R. F., & Waugh, R. F. (2011). *Applications of Rasch measurement in learning environments research*. Sense Publishers. https://doi.org/10.1007/978-94-6091-493-5

Chang, H. (2015). Psychometrics behind computerized adaptive testing. *Psychometrika, 80*(1), 1-20. https://doi.org/10.1007/s11336-014-9401-5

Chen, W. H., & Thissen, D. (1997). Local dependence indices for item pairs using item response theory. *Journal of Educational and Behavioral Statistics, 22*(3), 265-289. https://doi.org/10.3102/10769986022003265

Čisar, S. M., Čisar, P., & Pinter, R. (2016). Evaluation of knowledge in object oriented programming course with computer Adaptive tests. *Computers & Education, 92-93*, 142-160. https://doi.org/10.1016/j.compedu.2015.10.016

Čisar, S. M., Radosav, D., Markoski, B., Pinter, R., & Čisar, P. (2010). Computer adaptive testing of student knowledge. *Acta Polytechnica Hungarica, 7*(4), 139-152.

Costa, P., & Ferrão, M. E. (2015). On the complementarity of classical test theory and item response models: Item difficulty estimates and computerized adaptive testing. *Ensaio: Avaliação e Políticas Públicas em Educação* [*Essay: Evaluation and Public Policies in Education*]*, 23*(88), 593-610. https://doi.org/10.1590/S0104-40362015000300003

Cupani, M., Ghio, F., Leal, M., Giraudo, G., Castro Zamparella, T., Piumatti, G., Casalotti, A., Ramírez, J., Arranz, M., Farías, A., Padilla, N., & Barrionuevo, L. (2016). Desarrollo de un banco de ítems para medir conocimiento en estudiantes universitarios [Development of an item bank to measure knowledge in university students]. *Revista de Psicología* [*Psychology Journal*], *25*(2), 1-18. https://doi.org/10.5354/0719-0581.2017.44808

Cupani, M., Zamparella, T. C., Piumatti, G., & Vinculado G. (2017). Development of an item bank for the assessment of knowledge on biology in Argentine university students. *Journal of Applied Measurement, 18*(3), 360-369.

Doran, Y. J. (2017). The role of mathematics in physics: Building knowledge and describing the empirical world. *ONOMÁZEIN Número Especial LSF y TCL Sobre Educación y Conocimiento* [*ONOMÁZEIN Special Issue LSF and TCL on Education and Knowledge*], *13*(2), 209-226. https://doi.org/10.7764/onomazein.sfl.08

Dorans, N. J., & Kingston, N. M. (1985). The effects of violations of unidimensionality on the estimation of item and ability parameters and on item response theory equating of the GRE verbal scale. *Journal of Educational Measurement, 22*(4), 249-262. https://doi.org/10.1111/j.1745-3984.1985.tb01062.x

Downing, S. M., & Haladyna, T.M. (2006). *Handbook of test development*. Lawrence Erlbaum Associates.

Engelbrecht, J., Harding, A., & Du Preez, J. (2007). Long-term retention of basic mathematical knowledge and skills with engineering students. *European Journal of Engineering Education, 32*(6), 735-744. https://doi.org/10.1080/03043790701520792

Flores, A. H., & Gómez, A. (2009). Aprender matemática, haciendo matemática: La evaluación en el aula [Learning mathematics, doing mathematics: Assessment in the classroom]. *Educación Matemática* [*Mathematics Education*]*, 21*(2) 117-142.

Ghio, F. B., Cupani, M., Garrido, S. J., Azpilicueta, A. E., & Morán, V. E. (2019). Prueba para evaluar conocimiento en leyes: Análisis de los ítems mediante la aplicación del modelo de Rasch [Test to evaluate knowledge of law: Analysis of items applying the Rasch model]. *Revista Científica Digital de Psicología PSIQUEMAG* [*Digital Scientific Journal of Psychology PSIQUEMAG*]*, 8*(1), 105-116

Gierl, M. J., Bulut, O., Guo, Q., & Zhang, X. (2017). Developing, analyzing, and using distractors for multiple-choice tests in education: A comprehensive review. *Review of Educational Research, 87*(6), 1082-1116. https://doi.org/10.3102/0034654317726529

Haladyna, T. M., & Rodriguez, M. C. (2013). *Developing and validating test items*. Routledge. https://doi.org/10.4324/9780203850381

Han, K. (C.) T. (2018). Conducting simulation studies for computerized adaptive testing using SimulCAT: An instructional piece. *Journal of Educational Evaluation for Health Professions, 15*, 20. https://doi.org/10.3352/jeehp.2018.15.20

Karjanto, N., & Yong, S. T. (2013). Test anxiety in mathematics among early undergraduate students in a British university in Malaysia. *European Journal of Engineering Education, 38*(1), 11-37. https://doi.org/10.1080/03043797.2012.742867

Kaya, Z., & Tan, S. (2014). New trends of measurement and assessment in distance education. *Turkish Online Journal of Distance Education, 15*(1), 206-217. https://doi.org/10.17718/tojde.30398

Kingsbury, G. G., & Houser, R. L. (1999). Developing computerized adaptive tests for school children. In Drasgow, F., & Olson-Buchanan, J. B. (Eds.), *Innovations in computerized assessment* (pp. 93-116). Erlbaum.

Kolen, M. J., & Brennan, R. L. (2014). *Test equating, scaling, and linking. Methods and practices*. Springer. https://doi.org/10.1007/978-1-4939-0317-7

Linacre, J. M. (2000). Computer-adaptive testing: A methodology whose time has come. In S. Chea, U. Kang, & J. M. Linacre (Eds.), *Development of computerized middle school achievement test.* Komesa Press.

Lindquist, M., Philpot, R., Mullis, I. V. S., & Cotter, K. E. (2017). TIMSS 2019 mathematics framework. In I. V. S. Mullis, & M. O. Martin (Eds.), *TIMSS 2019 assessment frameworks* (pp. 11-25). TIMSS & PIRLS International Study Center, Boston College.

Messick, S. (1989). Meaning and values in test validation: The science and ethics of assessment. *Educational Researcher, 18*(2), 5-11. https://doi.org/10.3102/0013189X018002005

Navas, M. J. (1996). Equiparación de puntuaciones [Equalization of scores]. In J. Muñiz (Ed.), *Psicometría* [*Psychometry*] (pp. 293-370). Universitas, S. A.

Olea, J., Ponsoda, V., & Prieto, G. (1999). *Tests informatizados: Fundamentos y aplicaciones* [*Computerized tests: Fundamentals and applications*]. Pirámide.

Pallant, J., & Tennant, A. (2007). An introduction to the Rasch measurement model: An example using the hospital anxiety and depression scale (HADS). *British Journal of Clinical Psychology, 46*(1),1-18. https://doi.org/10.1348/014466506X96931014466506X96931

Phankokkruad, M. (2012). Association rules for data mining in item classification algorithm: Web service approach. In *Proceedings of the 2nd International Conference on Digital Information and Communication Technology and its Applications* (pp. 463-468). https://doi.org/10.1109/DICTAP.2012.6215408

Pollock, M. J. (2002). Introduction of CAA into a mathematics course for technology students to address a change in curriculum requirements. *International Journal of Technology and Design Education, 12*(3), 249-270. https://doi.org/10.1023/A:1020229330655

Programa Estado de la Nación. (2011). *Tercer informe estado de la educación* [*Third state of education report*]. PEN.

Putwain, D. W., Connors, L., & Symes, W. (2010). Do cognitive distortions mediate the test anxiety–examination performance relationship? *Educational Psychology, 30*(1), 11-26. https://doi.org/10.1080/01443410903328866

R Core Team. (2017). R: A language and environment for statistical computing. *R Foundation for Statistical Computing, Vienna, Austria*. https://www.R-project.org/

Reeve, B. B., Hays, R. D, Bjorner, J. B., Cook, K. F, Crane, P. K., Teresi, J. A., Thissen, D., Revicki, D. A., Weiss, D. J., Hambleton, R. K., Liu, H., Gershon, R., Reise, S. P., Lai, J. S., & Cella, D. (2007). Psychometric evaluation and calibration of health-related quality of life item banks: Plans for the patient-reported outcomes measurement information system (PROMIS). *Medical Care, 45*(5), S22-S31. https://doi.org/10.1097/01.mlr.0000250483.85507.04

Rodriguez, M. C. (2005). Three-options are optimal for multiple-choice items: A meta-analysis of 80 years of research. *Educational Measurement, 24*(2), 3-13. https://doi.org/10.1111/j.1745-3992.2005.00006.x

Rodríguez, P., Díaz, M., & Correa, A. (2015). Los aprendizajes al ingreso en un Centro Universitario Regional [Learning upon admission to a Regional University Center]. *Intercambios, 2*(1), 90–99. https://ojs.intercambios.cse.udelar.edu.uy/index.php/ic/article/view/47/149

Rojas, L., Mora, M., & Ordóñez, G. (2018). Asociación del razonamiento cuantitativo con el rendimiento académico en cursos introductorios de matemática de carreras STEM [Association of quantitative reasoning with academic performance in introductory mathematics courses of STEM careers]. *Revista Digital Matemática, Educación e Internet* [*Digital Journal of Mathematics, Education and the Internet*]*, 19*(1), 1-13. https://doi.org/10.18845/rdmei.v19i1.3851

Rojas-Torres, L., & Ordóñez, G. (2019). Proceso de construcción de pruebas educativas: El caso de la prueba de habilidades cuantitativas [Development process of educational tests: Quantitative ability test]. *Evaluar* [*Evaluate*]*, 19*(2), 15-29. https://doi.org/10.35670/1667-4545.v19.n2.25080

Smith, E. V. (2002). Detecting and evaluating the impact of multidimensionality using item fit statistics and principal component analysis of residuals. *Journal of Applied Measurement, 3*(2) 205-231.

Tennant, A., & Conaghan, P.G. (2007). The Rasch measurement model in rheumatology: What is it and why use it? When should it be applied and what should one look for in a Rasch paper? *Arthritis Care & Research, 57*(8), 1358-1362. https://doi.org/10.1002/art.23108

Tseng, W. (2016). Measuring English vocabulary size via computerized adaptive testing. *Computers & Education, 97*, 69-85. http://doi.org/10.1016/j.compedu.2016.02.018

Universidad Nacional de Córdoba. Secretaría de Asuntos Académicos. Programa de Estadística Universitaria (2020). *Anuario estadístico 2019* [*Statistical Yearbook 2019*]. http://www.interior.gob.es/web/archivos-y-documentacion/anuario-estadistico-de-2019

Vie, J. J., Popineau, F., Bruillard, E., & Bourda, Y. (2017). A review of recent advances in adaptive assessment. In A. Peña-Ayala (Ed.), *Learning analytics: Fundaments, applications, and trends. Studies in systems, decision and control*. Springer, Cham. https://doi.org/10.1007/978-3-319-52977-6_4

Wainer, H. (2000). *Computerized adaptive testing: A primer*. Lawrence Erlbaum Associates. https://doi.org/10.4324/9781410605931

Yen, W. M. (1993). Scaling performance assessments: Strategies for managing local item dependence. *Journal of Educational Measurement, 30*(3), 187-213. https://doi.org/10.1111/j.1745-3984.1993.tb00423.x

Zamora Araya, J. A. (2015). Análisis de la confiabilidad de los resultados de la prueba de diagnóstico matemática en la Universidad Nacional de Costa Rica utilizando el modelo de Rasch [Reliability analysis diagnostic mathematics test at the National University of Costa Rica]. *Actualidades en Psicología* [*News in Psychology*]*, 29*(119), 153-165. https://doi.org/10.15517/ap.v29i119.18693