Revista ELectrónica de Investigación y EValuación Educativa

RELIEVE

ISSN: 1134-4032

e-Journal of Educational Research, Assessment and Evaluation

Validity of the Automatic Generation of Ítems for the Basic Competences Exam (Excoba)

Validez del generador automático de ítems del examen de competencias básicas (Excoba)

Ferreyra, María Fabiana[(1)] & Backhoff Escudero, Eduardo[(2)]

(1) Universidad de Baja California, México (2) Universidad Nacional Autónoma de México

Resumen

La Generación Automática de Ítems (GAI) es el proceso con el cual se diseñan y elaboran reactivos de una prueba, así como versiones completas de exámenes conceptual y estadísticamente equivalentes. Los Generadores Automáticos de Ítems se desarrollan con el apoyo de sistemas informáticos, que los hacen sumamente eficientes. Con esta idea se creó el generador automático de reactivos GenerEx del Examen de Competencias Básicas (Excoba). Si bien la GAI representa un gran avance en el desarrollo de la evaluación psicológica y educativa, validar la gran cantidad de reactivos y exámenes que se generan de manera automática es un reto metodológico para la psicometría. Este trabajo tuvo el propósito de describir una propuesta para analizar la estructura interna y equivalencia psicométrica de los exámenes generados con el GenerEx, así como describir el tipo de resultados que se obtienen para lograr este propósito. La propuesta se fundamenta en la forma de seleccionar las muestras de reactivos, partiendo del principio de que los ítems y exámenes obtenidos deben ser equivalentes psicométricamente. El estudio se basa en tres tipos de análisis con marcos conceptuales diferentes y complementarios: la Teoría Clásica de los Test, la Teoría de Respuestas al Ítem y el Análisis Factorial Confirmatorio. Los resultados indican que el GenerEx produce exámenes psicométricamente similares, aunque con ciertos problemas en algunas áreas temáticas. La metodología permitió obtener una buena descripción del funcionamiento psicométrico del GenerEx y de la validez interna de dos versiones generadas al azar. Los análisis se pueden complementar con un estudio cualitativo de las deficiencias detectadas.

Fecha de recepción
25 de Octubre de 2015

Fecha de aprobación
26 de Octubre de 2015

Fecha de publicación
1 de Febrero de 2016

Palabras clave:

Generación Automática de Ítems, tests educativos, validez de constructo, estructura factorial, análisis de ítems

Abstract

Automatic Item Generation (AIG) is the process of designing and producing items for a test, as well as generating different versions of exams that are conceptually and statistically equivalent. Automatic Item Generation tools are developed with the assistance of information systems, which make these tools very efficient. Under this aim, GenerEx, an automatic item generation tool, was developed. GenerEx is used to automatically generate different versions of the Basic Competences Exam (Excoba). Even though AIG represents a great advance for the development of psychological and educational assessment, it is a methodological challenge to obtain evidence of validity of the enormous quantity of possible items and tests generated in an automatic process. This paper has the purpose of describing an approach to analyze the internal structure and the psychometric equivalence of exams generated by GenerEx and, additionally, to describe kinds of results obtained to reach this objective. The approach is based on the process for selecting samples from the generation tool, founded on the assumption that items and exams must be psychometrically equivalent. This work includes three kinds of conceptually different and complementary analysis: the Classical Test Theory, Item Response Theory and Confirmatory Factor Analysis. Results show that GenerEx produces psychometrically similar exams; however there are problems in some learning areas.

Reception Date
2015 October 25

Approval Date
2015 October 26

Publication Date:
2015 February 1

Autor de contacto / Corresponding author

Ferreyra, Maria Fabiana. Métrica Educativa, Alvarado 921, Zona Centro. Ensenada, Baja California, C.P. 22800 (México). fferreira@metrica.edu.mx

The methodology was useful for obtaining a description about GenerEx's psychometric functioning and the internal structure of two randomly generated versions of Excoba. Analysis can be complemented by a qualitative study of this item deficiencies.

Keywords:

Automatic Item Generation, Educational Testing, Construct Validity, Factor Structure, Item Analysis

Automatic Item Generation (AIG) refers to the process of designing and producing test items that are conceptually and statistically equivalent and which are developed with the aid of computer systems (Gierl & Lai, 2012). This procedure requires the participation of specialists who design item models, as well as complex statistical methods to validate the quality and equivalence of the items generated.

The conceptual origins of AIG can be found in the works of Hively, Patterson & Page (1968). These authors stated that items could be generated through item forms that contained explicit rules, with which it was possible to generate items that measured the same cognitive skills, but that did not necessarily offer the same psychometric properties, such as their level of difficulty and discrimination.

AIG development progressed with the emergence of cognitive methods for instruction and diagnostic assessment. These methods, however, focused on teaching and not on tests. As a result, cognitive models were developed, but psychometric implications, such as the equivalence between different tests, were not explored.

The third step was made when perspectives from psychometry and cognitive models were assimilated, which gave rise to two theoretical proposals: the Strong Theory (Embretson, 1999) and the Weak Theory (Bejar, 1993). The Strong Theory is based on cognitive task models, where aspects that affect the level of complexity (or difficulty) of the items generated are specified and used in accordance with the relevant theoretical framework. Each cognitive task model forms the basis for the creation of multiple item models[1], which, in turn, generate a range of equivalent items. According to Embretson (1999), it is possible to predict and control the psychometric properties of items with robust cognitive modeling. Gierl and Lai (2012) state that AIG based on the Strong Theory has seen little development in educational test design because it has focused mainly on basic psychological processes. Consequently, few cognitive theories have been developed that can be used as a basis for designing a range of question models to address the assessment needs of different areas of education.

On the other hand, the Weak Theory uses templates (or shells) to design item models that generate equivalent or isomorphic questions. A template is a kind of conceptual mold made up of a basic syntactical structure (this is a task that students must perform), with fixed and variable components, which, when complete with pre-established rules, allow the generation of a set of similar questions (Haladyna & Shindoll, 1989). For multiple-choice questions, an item model must include the following parts: the stem, the answer choices and supplementary information (Gierl & Lai, 2011). The stem contains the context, the content and the question that the examinee must answer. The options must include the correct answer and one or more distractors. The supplementary information includes any additional material necessary to generate the questions (texts, pictures, tables, diagrams). Both the basis for the item and the answer choices can be subdivided into components (sentences, words, letters, symbols, numbers, etc.). The items generated by a template are called siblings or instances. If the items generated with the item model measure content with similar levels of difficulty, the items are said to be isomorphic. In this case, item developers modify questions' superficial characteristics, which do not alter their difficulty, in order to produce the isomorphic items.

Although this proposal does not require a cognitive theory and is very appropriate for the range of educational exams, it does also have limitations. One limitation is that the

producers of item models must predict the psychometric properties of items so that they are similar. However, this is not always achieved. Another drawback is that sometimes, when producing isomorphic items, superficial changes are made in the templates, leading to items that are too similar or practically identical.

Although AIG has been around for over 40 years, one essential problem has hindered its progress: its validity. The simplest strategy for analyzing the answers to the sibling-items generated by AIG is to study each answer individually as an independent entity. However, if we consider that an item generator produces hundreds or thousands of items, this becomes an inefficient and monumental task. Therefore, alternative and innovative models are required to analyze the thousands of questions obtained by AIG.

Sinharay and Johnson (2012) described three models for analyzing and calibrating items produced by AIG. The first model involves predicting the psychometric properties of items, and their difficulty in particular, in accordance with the characteristics of task models used to generate the sibling-items. The second model considers the dependence between parameters from the same family of items. The third model is a combination of the previous two.

With regard to the first approach, researchers like Embretson (1999) and Holling, Bertling and Zeuch (2009) used the Linear Logistic Test Model (LLTM, as proposed by Fischer in 1973), which is an extension of the Rasch model. To do this, a cognitive model supporting all content, and therefore the items generated, is required. This means that the model is based on a Strong Theory.

The second model is based on the items from a template being grouped in families, with the aim of estimating the model parameters at the family level. The two most widely developed procedures are the Identical Siblings Model (ISM) and the Related Siblings Model (RSM). The ISM, by Hombo and Dresher (2001), takes on a one-answer-only function for all items from the same family. This model bears some limitations because it does not consider the variations within a family. Glass and Van der Linden (2003) proposed RSM with the aim of solving this problem by incorporating a connected structure between instances from the same family. RSM is applied mainly in adaptive tests, where examinees' ability plays a primordial role. This analysis focuses on the study of instances as isomorphic entities within the same family and not on the structure of the exam as a whole, with a set number of items.

The third model is a combination of LLTM and RSM approaches within another approach called the Linear Item Cloning Model (LICM), developed by Geerlings, Glas and Van der Linden (2011). The authors used a three-parameter normal ogive model to find the odds of answering an item correctly. For the reasons already stated, it is inferred that this methodology must also be used for strong-theory AIG.

Basic Competences Examination (Excoba)

The Basic Competences Examination (Excoba, a Spanish acronym) is a standardized, high-stake exam used to select students aspiring to enter Upper Secondary Education and Higher Education in Mexico (known in Mexico as 'EMS' and 'ES' respectively). This test originates from the Basic Knowledge and Skills Examination (Exhcoba), a large-scale, multiple-choice, computer-based exam (see Backhoff & Tirado, 1992; Backhoff, Ibarra & Rosas, 1995). Although Excoba retains the principle of its predecessor, insofar as it assesses basic and essential knowledge students acquire during the education, its structure and makeup is completely different and innovative.

The structure of Excoba is tied to the national curriculum, and as such, it assesses basic academic competences set out in the syllabi of obligatory education. It offers an

innovative approach with regard to the way in which these academic competences are assessed, as it distances itself from the multiple-choice format and is geared towards more 'authentic or natural' ways of assessing learning.

The version of Excoba used for admission into Upper Secondary Education (Excoba/MS) provides a measure of competences students are expected to have learned from national syllabi and study programs at the basic education level[2], using a fixed number of items: 120 in total (40 from elementary education and 80 from secondary education). Table 1 shows the type and number of competences assessed by this test. At elementary-school level, competences in Mathematics and Language are included, and at lower-secondary-school level, competences in Mathematics, Language, Natural Sciences (biology, physics and chemistry) and Social Sciences (history, geography and civics).

Table 1 - Number of questions for the basic academic competences that make up the Excoba/MS

Competences	Elementary	Lower Secondary	Total
Mathematics	20	20	40
Language	20	20	40
Natural Sciences	-	20	20
Social Sciences	--	20	20
Total	40	80	

NB: Mathematics competences at elementary level are called 'Mathematical skills' and at lower secondary level, 'Mathematics'. Language competences at elementary level are called 'Language skills' and at lower secondary level, Spanish.

As previously mentioned, the Excoba/MS questions are not multiple-choice, or at least not in the traditional sense, where the student must select an option from several given possibilities. The main types of items in Excoba are: (1) Constructed-Response: a numerical or algebraic solution is written literally; (2): Semi-Constructed-Response: graphic or conceptual elements are placed or moved on maps, graphs, diagrams, plans or formats (e.g.: by finding geographic coordinates on a plan); and (3): multiple-multiple-choice: the answer is given by selecting three or more options (e.g. by selecting elements that make up a category).

Some of these Excoba items are graded dichotomically (right-wrong). This is the case for constructed-response items. Others are graded following the partial-credit model, according to the number of answers required for each item; this is the case for semi-constructed and multiple-multiple-choice items. The highest score that can be awarded for each item is one unit. For the partial-credit model, the unit is divided by the number of parts the item contains; for example, if the item asks the candidate to place four fractions on a number line, each correctly-placed fraction is worth 0.25 points.

The Excoba Automatic Item Generator

In order to produce sibling-items for the Excoba, the Automatic Exam Generator (GenerEx, a Spanish acronym) was developed in line with the assessment needs for the varying content of the curriculum at the basic education level. As a result, GenerEx belongs to the generators of the Weak Theory, as it is not based on any particular cognitive theory that may give a detailed explanation of the cognitive processes of each academic competence being assessed.

As mentioned, AIG requires the creation of item models that provide, at the least, the following elements: 1) a definition of the competence being assessed; 2) a specific strategy to assess this competence; and 3) a template, with the rules and elements (either conceptual or graphic). These templates enable the creation of a family of items containing at least one parent-item, from which sibling-items are derived. The core idea is that the item model must be able to automatically generate a large number of instances with which students can be consistently assessed for a given academic competence. In order to achieve this, the sibling-items must have equivalent metric and conceptual properties.

Fraction Representationn competence with one parent-item (<="" i="">). From this parent-item, several sibling-items are derived, made up of different fractions of different shapes (square, pentagon and triangle).

In each item model, the context in which the question is presented, the action that the student must take to answer the question and the rules used to generate each sibling-items automatically must all be given. In the previous example, all the sibling-items ask the student to select the parts of a geometric figure shown by a specific fraction. This item model allows several shapes to be selected, and in turn, for each shape, several fractions can be chosen. This selection may be random or fixed, as desired.

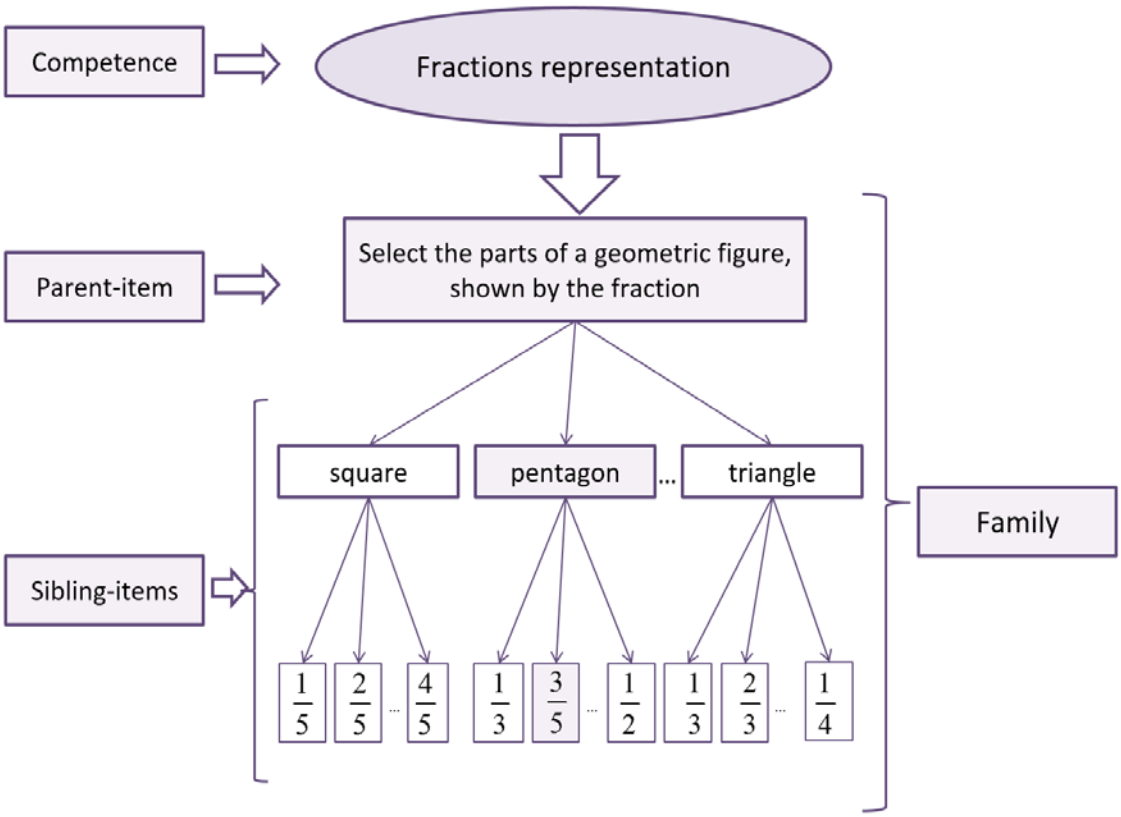

Figure 1. Family of questions for the Fraction Representation competence, from the Mathematics section

Figure 2 gives a visual display of an instance that can be produced with GenerEx. In this case, the question shows a pentagon divided into equal parts and the student is asked to select the parts that represent the fraction 3/5.

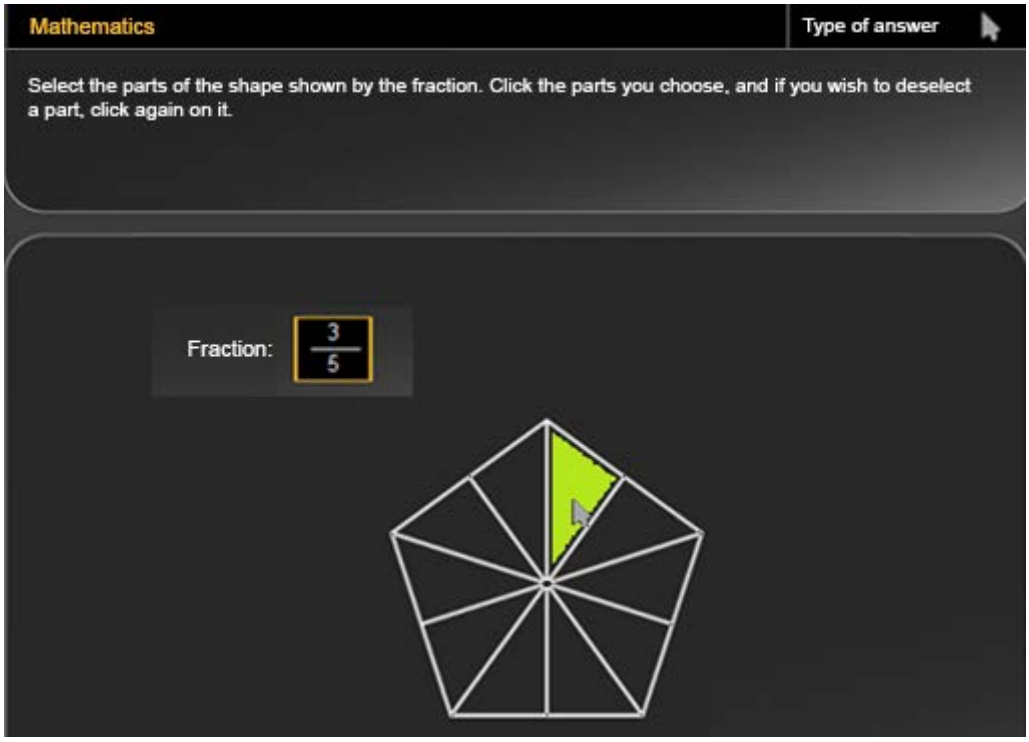

Figure 2. An instance of a family of items generated for the Fraction Representation competence, from the Mathematics section

Problem Statement

Automatic Item Generators like GenerEx produce tens or hundreds of conceptually equivalent questions, which opens up the possibility of creating hundreds or thousands of parallel exams. If the question models are well-designed, they will have equivalent psychometric properties, and the exams created with these questions will display similar internal structures.

As previously mentioned, AIG poses new challenges for psychometrics, and certainly the most important of these is how its validity can be ensured. This is because it would be impossible to become familiar with the psychometric properties of all the sibling-items that it is possible to generate with GenerEx, and this is even more the case for the internal structure of all the possible versions that can be created with the combination of 120 item models in the case of Upper Secondary Education.

Consequently, this paper aims to propose a way of studying the validity of GenerEx and show examples of results obtained using this methodology. More specifically, we set out to provide evidence of the validity of this generator at two levels: for the exams generated and for each of the question models (with which the items are created).

Method

The methodological approach to studying the validity of GenerEx involved making a series of comparative studies of the items, at three levels: 1) the exam level, with the aim of looking into the measures of central tendency, the dispersion and the reliability of the tests, in addition to comparing different versions of the exam and studying the behavior of the six subject areas (that make up the exam); 2) the family level of items, in order to analyze and compare items from the same competence, study their psychometric properties and observe whether they are grouped around the corresponding construct or latent trait; and (3) the component level, where components of items are studied so as to make a decision about their quality.

This was achieved with tools proposed by the Classical Test Theory (CTT), Item

Response Theory (IRT) and Confirmatory Factorial Analysis (CFA). In particular, the analyses based on IRT were made with Rasch's classical model for dichotomous data (1961) and with the Rasch model for partial credit items (Masters, 1982).

Given the space it would take to address the three types of analysis, this paper is limited to the first level, which deals with the general validity of the exam.

Sample of questions and students

Two parallel exams, made up of six subject areas and 120 questions, were generated for the comparative study of GenerEx. These will be called version A (VA) and version B (VB) for this study. The questions for both versions of the Excoba/MS were obtained randomly, and insofar as was possible, an attempt was made to ensure that items belonging to the same family of questions did not focus on the same aspects.

Figure 3 shows the general makeup of the exam for admission to upper secondary school. In both versions the makeup was the same, but had different siblings for each competence. In this figure it can be noticed that Excoba/MS measures two major competences learned at elementary level (Language skills and Mathematical skills) and four from secondary level (Spanish, Mathematics, Natural Sciences and Social Sciences). Each competence is made up of 20 question models (for example, HV01, HV02…HV20) [3] and each question model can generate a large number of conceptually equivalent items (e.g. from HV1 it is possible to derive I1, I2…In). For this study, only one item per model was generated, which formed an exam with 120 items (20 items per area).

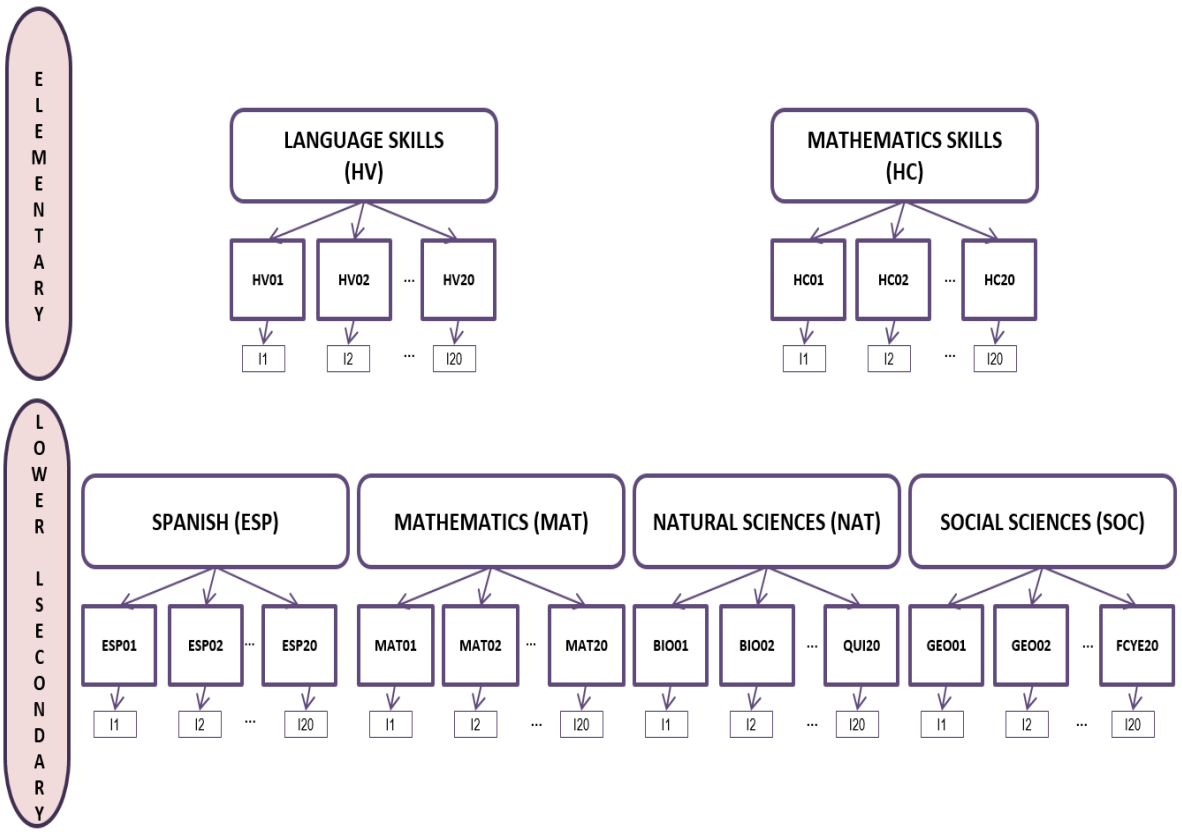

Figure 3. Makeup of the Excoba/MS

Both versions of the exam were given to groups of students aspiring to enter the Lázaro Cárdenas Federal High School (abbreviated to PFLC in Spanish), located in the city of Tijuana, Baja California. The PFLC was founded in 1946 as the first upper secondary-

level institution in Tijuana. There are currently over 4,000 students enrolled and it is one of the most prestigious schools in the state, with the best results in the National Assessment of Academic Achievement in Schools (Evaluación Nacional del Logro Académico en Centros Escolares, or ENLACE in Spanish), as attested by the school's website (http://dir.lazarocardenas.edu.mx/). 401 students answered VA, whereas 299 answered VB. The students were between 15 and 16 years of age, and 59% were female and 41% male. The students' grade point average in secondary education, on a scale of 5 to 10, was 9.13 with a standard deviation of 0.57. Participation was on a voluntary basis and encouraged by the institution. The results were given back to the institution to help better prepare future students. A random selection was made to determine which of the two versions of the exam each student would take.

Results analysis

Versions A and B of the Excoba/MS were analyzed and compared on two levels.

1) For the full exam (made up of 120 questions), the following calculations were made: frequency distribution and normality; measures of central tendency and dispersion; bias and kurtosis. Similarly, in order to analyze the internal consistency, the point-biserial correlation indices and Cronbach's alpha coefficient were obtained. With the Rasch model, the fit, item-measure correlation and discrimination indices were calculated, in addition to the Wright map.

2) For each of the six subject areas (20 items) that make up the exam, the following indicators were obtained: difficulty, point-biserial correlation and reliability; measure, level of fit (internal or external), point-measure correlation and discrimination; in addition to the indices and factor loading for item clustering for each area of the test.

The statistical analysis was made with the help of the following programs: SPSS 17.0 (SPSS, 2008), Winsteps (Linacre, 2010) and EQS 6.1 (Bentler, 2006). Table 2 shows the criteria and limits set to assess the quality of individual items and the exam as a whole, according to the psychometric model used. For example, the minimum acceptable point-biserial correlation for the questions was set at 0.2; the minimum reliability (a) for the subject areas of the exam (with 20 items) had to be at least 0.6, whereas the reliability of the test as a whole (120 questions) had to be equal to or greater than 0.9.

Table 2 - Criteria used for the statistical analysis of items from the sample Excoba/MS

Psychometric models	**Statistics**	**Number of variables**	**Criterion**	
			Acceptable	Good
Classical Test Theory	Point-Biserial correlation		≥ 0.2	
	Alpha (a)	20	≥ 0.6	
		120	≥ 0.9	
Item Response Theory	Point-measure correlation		≥ 0.2	
	Infit-Outfit MNSQ		≥ 0.8 y ≤ 1.3	
	Discrimination		≥ 0.8	
Confirmatory Factorial Analysis	Factor loading		≥ 0.20	≥ 0.30
	c^2		≥ 0.01	≥ 0.05
	NNFI		≥ 0.90	≥ 0.95
	CFI		≥ 0.90	≥ 0.95
	RMSEA		< 0.08	< 0.05

Results

The total number of Excoba/MS items assessed was 117, out of the original 120. This loss was due to three item models being discarded: one from language (HV19) and another from history (HIS06), both of which had a design problem, and the remaining model, from mathematics (MAT14), was rejected due to the difficulties in decoding the students' answers.

Full versions of the exam

The metric properties of both full versions of the Excoba/MS are described and compared below. First of all, the measures of central tendency, dispersion, normality and reliability are presented, followed by the Wright map (which shows how students' abilities are on an equal level to the difficulty of the questions), the percent variance in each test, the fit indices, as well as the point-measure correlation and discrimination.

Table 3 shows how both versions have very similar indicators.[4] The mean values for the correct answers (levels of difficulty) are similar: for VA it is 60.9 (p = 0.52) and for VB, 58.1 (p = 0.50); in each version the dispersion is practically the same, as is their symmetry. With regard to the kurtosis, the distribution in VA is slightly leptokurtic (0.18), whereas VB is more platykurtic (-0.25). However, measurement errors are great, which means that differences between these values are not significant. Therefore, the data may indicate that the distributions are normal. Furthermore, the average for the point-biserial correlation of the questions was 0.26 for VA and 0.25 for VB. For both versions, the reliability is the same (a = 0.90), which strengthens the assumption that both versions are similar.

Table 3 - Central tendency, dispersion, normality and reliability indicators for VA and VB of the Excoba/MS

Indicator	Version A	Version B
N	163	119
Mean	60.9	58.1
Standard Deviation	10.6	10.4
Range	35 - 95	35 - 90
Symmetry	0.31	0.33
Kurtosis	0.18	-0.25
Point-biserial correlation	0.26	0.25
Reliability	0.90	0.90

Figure 4 shows the Wright map of the two full versions of the Excoba/MS. This map gives the distributions of the level of difficulty of the questions versus the students' abilities. For both versions, it can be noticed that the mean value of the difficulty of the items is greater (almost to one standard deviation) than the mean of the examinees' abilities. The most difficult subject area was mathematics at secondary level (MAT); the easiest was Spanish, also at secondary level (ESP), followed by Language Skills at elementary level (HV). In both exams, Mathematical Skills (HC) covered a range of levels of difficulty that ran from -2 to 3 logits, whereas Social Sciences (GEO, HIS, FCYE) and Natural Sciences (BIO, FIS, QUI) were within the -1 to 1 range. In general, although there are questions that exceed the students' abilities, for most of the items' difficulty levels are on an equal level with students' skills.

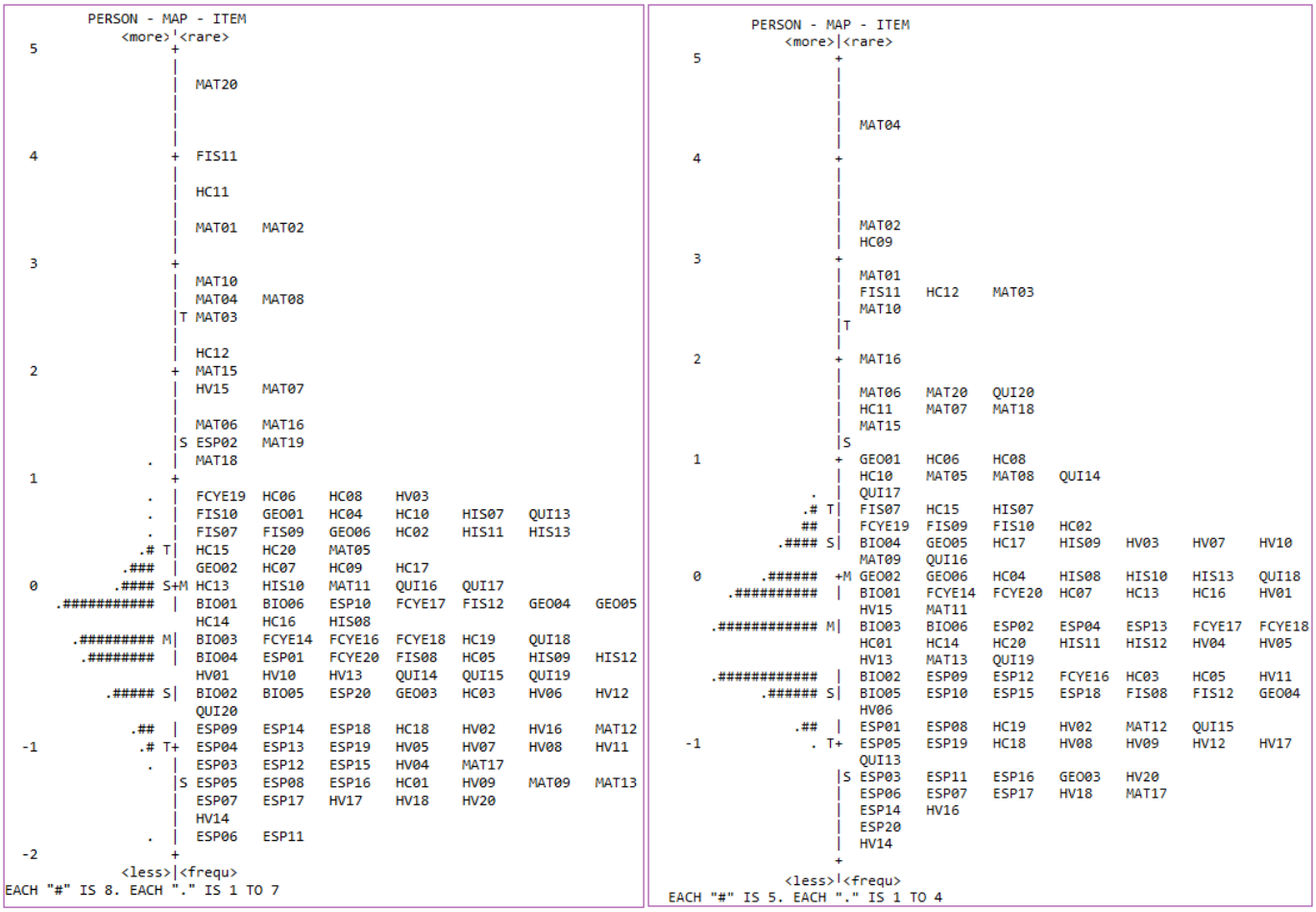

Figure 4. Wright maps of versions A (left) and B (right) of the Excoba/MS

Based on the Rasch analysis, no serious imbalances were found in any item; this indicates that there were no problems related to randomization or determinism, neither close to nor far from the measurement area of each item. However, some failures in correlation or discrimination were detected for questions HV15 and HC07 in both versions; in version A similar problems were found in questions ESP01, QUI13 and QUI16, whereas in version B the problem questions were HC09, ESP02, MAT07, QUI14 and QUI20.

Finally, the measurements (students' ability and the level of difficulty of the items) explain 38.5% and 37.3% of the variance for versions A and B respectively. The average point-measure correlation index was 0.29 for version A and 0.28 for version B. These results indicate acceptable values and are very similar for two exams that represent the same latent trait or construct.

Comparison by subject area in the exam

Table 4 shows the psychometric behavior of the six areas of both versions of the exam. Values outside of the desired range are given in bold type. In both versions, the areas with greater and lesser difficulty are those areas related to mathematics and language respectively. On the other hand, all subject areas displayed very few differences in the levels of difficulty, the greatest being 0.05, in Language Skills.

Table 4 - Results of the analysis with the CTT of the Excoba/MS for VA and VB, by subject areas

Subject area	k	Version A				Version B			
		n	p	ptbis	α	n	p	ptbis	α
Language Skills	19	289	0.68	0.24	0.633	189	0.63	0.21	**0.547**
Mathematical Skills	20	396	0.38	0.34	0.784	301	0.38	0.35	0.788
Spanish	20	290	0.69	0.21	**0.587**	270	0.66	0.30	0.706
Mathematics	19	396	0.26	0.26	0.655	296	0.24	0.27	0.691
Natural Sciences	20	289	0.48	0.22	0.612	216	0.47	**0.16**	**0.502**
Social Sciences	19	397	0.47	0.48	0.869	298	0.48	0.48	0.877

NB: k = number of items, n = sample size, p = average difficulty, Ptbis = point-biserial correlation, α = Cronbach's alpha. Values outside of the desired range in bold type.

With regard to the power of discrimination, measured by the point-biserial correlation index (ptbis) of each question, similar average scores are observed for most subject areas, with the exception of Spanish and Natural Sciences. In VB of this latter subject area, the required minimum of 0.20 was not reached. Finally, there are significant variations between both versions in the reliability of subject areas, measured by Cronbach's alpha, as is the case in Language skills (0.09 points), Spanish (0.11 points) and natural sciences (0.11 points). It is also important to note that the most reliable subject areas were Social Sciences and Mathematical skills.

Table 5 shows the results of the analysis performed with the Rasch model. By comparing the two versions from left to right, it is possible to notice how the amount of explained variance (Var) varies from one subject area to another. The areas with greatest divergence in the amount of variance explained through the measurements (ability and difficulty) are Mathematics (a difference of 18.2 points), Spanish (11.7 points) and Language skills (8.7 points); those with the least difference are Social Sciences (0.5) and Mathematical skills (0.7 points). On the other hand, the averages of the point-measure correlations of the different subject areas are very similar and in some cases equal, with Spanish being the area with the greatest difference (0.06).

Table 5 - Results of the Rasch model analysis of the Excoba/MS for versions A and B, by subject areas

Subject Area	k	Version A						Version B					
		n	Var(%)	*Pmed*	Problems			n	Var	*Pmed*	Problems		
					in	out	C/D				in	out	C/D
Lenguaje Skills	19	399	38.5	0.35			HV15	298	47.2	0.33			HV15
Mathematical Skills	20	401	32.9	0.42	HC07	HC07	HC07	301	32.2	0.42		HC07	HC09
Espanish	20	398	39.0	0.33				297	27.3	0.39			
Mathematics*	19	400	75.9	0.35		M9,10,12		300	57.7	0.37			M18
Natural Sciences	20	380	27.6	0.37		FIS11		273	34.5	0.34			
Social Sciences	19	397	38.6	0.50				298	39.1	0.50	G04	G04	

NB: (*) in VB only 18 items could be analyzed, because there were no correct answers for MAT19

k = number de items, n = simple size, Var = percent explained variance, Pmed = Point-measure correlation average, in = infit, out = outfit, C/D = point-measure correlation and/or

discrimination index. HC = mathematical skills, HV = language skills, M = Mathematics, FIS = Physics, G = Geography

For both versions of the Excoba, there were eight items with a fit problem (infit or outfit), out of a total of 117 questions. Questions HV15 and HC07 displayed poor behavior in both versions of the exam. For the rest – HC09, MAT09, MAT10, MAT12, FIS11 and GEO04 – this only happened in one of the tests analysis. Most problems were linked to high values of the outfit indicator. The two infit situations recorded were also due to exceeding the range of fit. These values indicate too much randomization far from the item's measurement area, in the former case, and too close to the measurement area in the latter case.

In order to understand the clustering of questions in each of the subject areas, confirmatory factorial analyses were performed for both versions of the Excoba/MS and the best-fitting models were sought. Table 6 shows, for the six subject areas, the indicators of fit and the number of factors identified for each case. As shown in the table, in all cases, the clustering models display good indicators of fit, the values of which are very similar between VA and VB. In Mathematical skills, Spanish and Social Sciences, only one factor was identified, whereas in Language skills, Mathematics and Natural Sciences, two factors were found.

Table 6 - Item clustering models for the six subject areas of the Excoba/MS, for VA and VB

Fit	Language Skills		Mathematical Skills		Mathematics		Spanish		Natural Sciences		Social Sciences	
	VA	VB	VA	VB	VA	VB	VA	VB	VA	VB	VA	VB
P	0.79	0.56	0.00	0.03	0.39	0.31	0.08	0.06	0.81	0.09	0.02	0.00
NNFI	1.05	1.02	0.93	0.94	0.98	0.98	0.96	0.91	1.07	0.81	0.97	0.95
CFI	1.00	1.00	0.94	0.95	0.98	0.98	0.97	0.95	1.00	0.84	0.98	0.97
RMSEA	0.00	0.00	0.03	0.03	0.01	0.02	0.02	0.03	0.00	0.03	0.03	0.04
Number of factors	2		1		2		1		2		1	
	HV15		HC07				ESP02		BIO01 y QUI19			
Items with no significant load	HV02 HV08 HV16	HV03 HV05	HC09		MAT16 MAT19	MAT07	ESP01 ESP18		QUI13 QUI16	BIO04 QUI14 QUI18 QUI20		

Both of the factors for Language skills involve reading and understanding texts, in addition to grammar and spelling. For Mathematics, the factors were linked, on the one hand, to number sense, algebraic thinking and information handling and, on the other hand, to shape, space and measurement. Finally, in Natural Sciences, there were factors linked on the one hand to biology and chemistry, and on the other hand, to physics.

Furthermore, in five of the six subject areas analyzed, questions were found with no significant loads in the respective model. The following items were detected for both versions: HV15, HC07, ESP02, BIO01 and QUI19. Language skills and Natural Sciences were the most divergent subject areas, as there were two or more problem items in each version.

As an example of how the items' factor loadings are distributed in one of the exam's subject areas, it is possible to take the two versions of the Mathematical skills section. Figure 5 gives the 20 items in each version and

shows both the similarity in factor loadings and the questions with loadings that are insufficient by our criteria (under 0.2). HC07 (in both versions) and HC09 in VB fall under this category.

A qualitative analysis of the item models in this subject area revealed that HC07 appealed to students' memory and recognition (recognize parts marked on a circumference); this is different to the rest of the skills in mathematics at elementary level, because most require understanding and application. For HC09, the two items were compared and although both require the student to calculate the area of a triangle, in one exercise the height of the side is shown inside the figure (VA) and in the other (VB), outside of the figure. It could be speculated that the latter picture may be confusing and could even disorientate students in their calculations.

Another important remark is that 14 items in VA and 16 in VB have factor loadings above 0.30. These results show us that both of the automatically-generated tests display very similar behavior, which lends weight to the validity of the exams obtained by GenerEx.

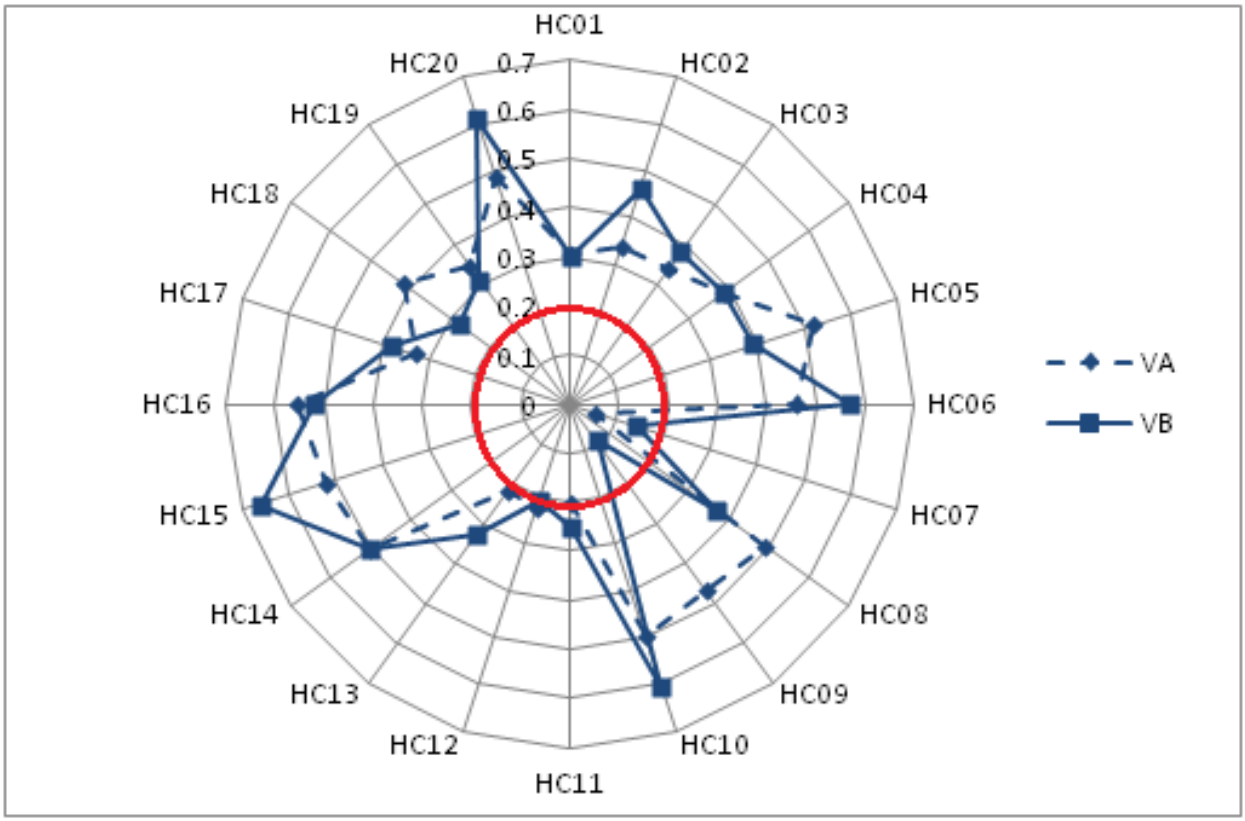

Figure 5. Distribution of factor loadings for Mathematical skills in the two versions of the Excoba/MS

Discussion and conclusions

Automatic item generators represent significant progress in psychological and educational assessment, as they allow the design and creation of a considerable number of conceptually and psychometrically equivalent questions (Gierl & Haladyna, 2012). AIG solve the problem of periodically having to produce unique tests, which quickly wear out when used on a massive and intensive scale, as is the case with entry exams. AIG development has undergone several stages, from AIG based on item template design, up to models based on cognitive conceptual frameworks (Haladyna,

2012). Without a doubt, new statistical tools and computer-aided assessment will be a strong driving force in consolidating AIG in the near future.

For the time being, the use of AIG for practical purposes imposes unprecedented challenges on psychometrics, owing to the need to find efficient and economical solutions to the difficulty of having valid evidence of automatically-produced questions and exams. It would not be feasible to envisage a validation process for each of the hundreds of questions generated and thousands of exams created. Consequently, we set ourselves the task of finding a method that responds to this problem, taking GenerEx, which is based on a Weak Theory (Gierl, Zhou & Alvez, 2008), as a reference. This was because it is not based on a task model that specifies the cognitive structures of the academic skills it assesses, but rather, it is based on knowledge that students are expected to have acquired in accordance with the Mexican curriculum (Ferreyra, 2014; Pérez-Morán, 2014).

This work rests on the assumption that item models, which define families from which sibling-items are derived and form a version of the exam, structure the assessment of each competence. In principle, each model specifies and controls both the competence being assessed and the difficulty of the questions generated, in such a way that assessment tasks use similar criteria in assessing the same competence (Bejar, 2002; Gierl & Lai, 2011; Gierl, Zhou & Alves, 2008).

The proposed method considered, basically, three levels of analysis: at a macro-level, that of the exams generated; at a meso-level, that of the families of questions; and at a micro-level, that of the sibling-items and their components. With regard to the validity of the exam as a whole (which was the aim of this study), the core idea of this methodological proposal was to compare the psychometric equivalence of two parallel versions of GenerEx, each of which was made up of 120 randomly-generated questions and did not share any components with the other one, and also the equivalence between subject areas that made up each exam. These psychometric comparisons were made by means of three complementary methodological approaches, which were based on the Classical Test Theory, the Item Response Theory (with the Rasch model) and the Confirmatory Factorial Analysis.

Subject areas (six in total) were also analyzed in both versions of the exam, where the mean value of the levels of difficulty and the point-biserial correlations of the Excoba/MS were calculated. In that manner, the basic psychometric indicators were obtained for each subject area based on the IRT. Finally, the factorial correspondence analyses were performed for both versions on the six subject areas of the Excoba, in order to compare the item clustering models and the factor loading of each question in the respective models.

With the three types of analysis performed, it was also possible to identify, for both versions of the exam and for the six subject areas, the items that displayed behavior outside of the acceptable ranges, in accordance with the criteria defined in this study (see table 2).

In sum, the methodology developed provided a good description of how GenerEx operates and the internal validity of two randomly-generated versions with basic statistical tools. The results can be well complemented with a qualitative analysis of the problems detected. This item generator produces psychometrically similar exams and questions, although it also displays problems in some subject areas and for certain specific items. Generally speaking, the shortcomings detected were found in both versions of the exam, although some imbalances were also identified between the two test versions. This points to a problem concerning the definition of the content that is fed into GenerEx, which will require a more detailed conceptual study.

Lastly, it must be said that the validation process of any measuring instrument must be

permanent, and that AIG development is still breaking ground across the world. As a result, a very interesting and fertile field of research, regarding methods of internal validation of automatic item generators, is on the horizon.

References

Backhoff, E. & Tirado, F. (1992). Desarrollo del Examen de Habilidades y Conocimientos Básicos. *Revista de la Educación Superior, 21* (3), 95-118. Retrieved from http://www.metrica.edu.mx/fileadmin/user_upload/pdf/1992_Desarrollo_del_EXHCOBA.pdf

Backhoff, E., Ibarra, M. & Rosas, M. (1995). Sistema Computarizado de Exámenes (SICODEX). *Revista Mexicana de Psicología, 12* (1), 55-62.

Bejar, I. I. (1993). A generative approach to psychological and educational measurement. In N. Frederikson, R. J. Mislevy & I. I. Bejar (Eds.). *Test theory for a new generation of tests* (pp. 323-359). Mahwah, NJ: Erlbaum.

Bejar, I. I. (2002). Generative testing: From conception to implementation. In S.H. Irvine & P.C. Kyllonen (Eds.), *Item generation for test development* (pp. 199-217). Mahwah, NY: Erlbaum.

Bentler, P. M. (2006). *EQS 6 Structural Equations Program Manual*. Encino, CA: Multivariate Software, Inc.

Embretson, S. E. (1999). Generating items during testing: psychometric issues and models. *Psychometrika, 64* (4) 407-433. doi: http://dx.doi.org/10.1007/BF02294564

Ferreyra M. F. (2014). *Metodología para analizar la estructura interna de un generador automático de reactivos* (Unpublished doctoral dissertation). Universidad Autónoma de Baja California, Ensenada, México.

Geerlings, H., Glass, C. A. W. & van der Linden, W. J. (2011). Modeling rule-based item generation. *Psychometrika, 76* (2), 337-359. doi: http://dx.doi.org/10.1007/s11336-011-9204-x

Gierl, M. J. & Haladyna, T. M. (2012). Automatic item generation: an introduction. In M. J. Gierl & T. M. Haladyna (Eds.), *Automatic item generation: Theory and practice* (pp. 3-12). Nueva York: Routledge.

Gierl, M. J. & Lai, H. (April, 2011). The Role of Item Models in Automatic Item Generation. Paper Presented at the *Annual Meeting of the National Council on Measurement in Education.* New Orleans, LA.

Gierl, M. J. & Lai, H. (2012). Using weak and theory to create item models for Automatic Item Generation: some practical guidelines with examples. In M. J. Gierl & T. M. Haladyna (Eds.). Automatic Item Generation: Theory and Practice. Nueva York: Routledge.

Gierl, M. J., Zhou, J. & Alves, C. (2008). Developing a Taxonomy of Item Model Types to Promote Assessment Engineering. *The Journal of Technology, Learning, and Assessment, (7)* 2.

Glas, C. A. W. & van der Linden, W. J. (2003). Computarized adaptive Testing with item cloning. *Applied Psychological Measurement, 27*, 247-261. doi: http://dx.doi.org/10.1177/0146621603254291

Haladyna, T. M. (2012). Automatic item generation: A historical perspective. In M. J. Gierl & T. M. Haladyna (Eds.), *Automatic item generation*: *Theory and practice* (pp. 13-25). Nueva York: Routledge.

Haladyna, T. M. & Shindoll, R. R. (1989). Shells: A method for writing effective multiple-choice test items. *Evaluation and the Health Professions, 12*, 97-104. doi: http://dx.doi.org/10.1177/016327878901200106

Hively, W., Patterson, H. L. & Page, S. H. (1968). A "universe-defined" system for arithmetic achievement tests. *Journal of Educational Measurement, 5*, 275-290. doi: http://dx.doi.org/10.1111/j.1745-3984.1968.tb00639.x

Holling, H., Bertling, J. P. & Zeuch, N. (2009). Automatic item Generation for probability word problems. *Studies in Educational Evaluation, 35*, 71-76. doi: http://dx.doi.org/10.1016/j.stueduc.2009.10.004

Hombo, C. & Dresher, A. (2001). *A simulation study of the impact of automatic item generation under NAEP-like data conditions*. Paper presented at the annual meeting of the National Council on Measurement in Education, Seatle, Wa, EE. UU.

Linacre, J.M. (2010). *Winsteps® (Version 3.70.0.2)* [Computer Software]. Beaverton, Oregon: Winsteps.com

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika, 47* (2), 149-174. doi: http://dx.doi.org/10.1007/BF02296272

Pérez-Morán, J. C. (2014). *Análisis del aspecto sustantivo de la validez de constructo de una prueba de habilidades cuantitativas* (Unpublished doctoral dissertation). Universidad Autónoma de Baja California, Ensenada, México.

Rasch, G. (1961). On General Laws and the Meaning of Measurement in Psychology. Proceedings of the *Fourth Berkeley Symposium on Mathematical Statistics and Probability, Volume 4: Contributions to Biology and Problems of Medicine*, 321-333. University of California Press: Berkeley, CA. Retrieved from http://projecteuclid.org/euclid.bsmsp/1200512895

Sinharay, S. & Johnson, M. (2012). Statistical modeling of Automatic Item Generation. In M. J Gierl & T. M. Haladyna (Eds.). *Automatic Item Generation: Theory and Practice*. N. Y., New York: Routledge.

SPSS Inc. (2008). *SPSS Statistics for Windows, Version 17.0.* Chicago: SPSS Inc.

Notes

[1] The terms *item model* and *template* are equivalent, for the purposes of this article. They must not be confused with *task model*, which is the cognitive model that underlies the trait being assessed

[2] TN: In Mexico, basic education (*educación básica*) refers to education at elementary school and lower secondary school.

[3] For historical reasons, the 'Language skills' questions are abbreviated HV, for 'Verbal Skills' (*Habilidades verbales*, in Spanish), and the 'Mathematical skills' questions are simplified to HC, for 'Quantitative Skills' (*Habilidades cuantitativas*, in Spanish). Verbal skills and quantitative skills are the names given to the elementary-level subject areas in the EXHCOBA examination

[4] It should be noted that the number of students considered for this analysis was considerably lower, because only those students who obtained a grade in all items were included

Acknowledgements

This study is part of the doctoral research carried out by María Fabiana Ferreyra at the *Universidad Autónoma de Baja California*, thanks to financing by Conacyt-México (Record number 247008)

Authors / Autores | **To know more / Saber más**

Ferreyra, Maria Fabiana (fferreyra@metrica.edu.mx).
Mathematics teacher from the *Instituto Nacional Superior del Profesorado Joaquín V. González*, Buenos Aires, Argentina. She holds a master's degree in Education Sciences and a Ph.D. in Education Sciences, both of which are from the Institute for Education Development and Research, part of the *Universidad Autónoma de Baja California*, Mexico. Her area of interest is the development and validation of large-scale learning tests, and teaching mathematics. She is currently a research associate at Métrica Educativa, A.C., Mexico. Her postal address is: Métrica Educativa, Alvarado 921, Zona Centro. Ensenada, Baja California, C.P. 22800 (México)

Backhoff-Escudero, Eduardo (ebackoff@gmail.com).
He holds a bachelor's degree in Psychology from the *Universidad Nacional Autónoma de México*, a master's degree in Education from the University of Washington and a Ph.D. in Education from the *Universidad Autónoma de Aguascalientes*. His area of interest is the development and validation of large-scale learning tests and computer-aided assessment. He has been Director of Tests and Measuring at the National Institute for Education Evaluation (INEE) in Mexico. He is currently a Member of the Governing Board of INEE

Revista **EL**ectrónica de **I**nvestigación y **EV**aluación **E**ducativa
E-Journal of Educational Research, Assessment and Evaluation

[ISSN: 1134-4032]