US-China Education Review B, July-Aug. 2022, Vol. 12, No. 4, 119-129
doi: 10.17265/2161-6248/2022.04.006

Math Test Items' Validation for Admission to Higher Education

Yurico Dulce Teresa Rivera Fernández, Omar Cuevas Salazar, Maria Esther Velarde Flores
Instituto Tecnológico de Sonora, Campus Náinari, Ciudad Obregón, México
Eduardo Guzmán de los Riscos
Universidad de Málaga, Málaga, España
Daniel Mocencahua Mora
Benemérita Universidad Autónoma de Puebla, Puebla, México

This study presents the validation of a math test items, whose function is to diagnose students in their admission to higher education. The test consists of 22 multiple choice items divided into seven main units. It was applied to a population of 940 students. Methodology comprised of three phases. The first one was the validation by experts, one from mathematics field and another expert in development of computerized adaptative tests. The second phase was the pilot testing. The third phase consisted of the process to get the validation and reliability by analyzing items' difficulty and discrimination levels through a biserial correlation. Reliability was proved through a Rasch model developed with the support of R and SPSS software. The findings presented a positive difficulty asymmetry of 0.59, a discrimination index of 0.48, and a test reliability of 0.74. So, it is established that the test is suitable to be applied.

Keywords: admission tests, higher education, item validation, mathematics, Rasch model

Introduction

Higher education institutions usually focus their efforts in direction of constant improvement in order to achieve goals and objectives of educational standards established by assesment organisms. So then, evaluation is a way to establish quality assurance and to identify opportunity areas to work in the achievement of continuos improvement. Therefore, it is necessary to set certain tests to assess each one of the teaching-learning process areas.

For a test to be correctly established, it is requerired to run a validation process to confirm the convenience of its development and to prove that it was correctly optimized and standarized for an specific purpose. According to the Standards for Educational and Psychological Testing, validity refers to the level in which evidence and

Yurico Dulce Teresa Rivera Fernández, Ph.D., Professor, Instituto Tecnológico de Sonora, Campus Náinari, Antonio Caso 2266, Villa Itson, 85130, Ciudad Obregón, Sonora, México.
Eduardo Guzmán de los Riscos, Ph.D., Professor, Dpto. de Lenguajes y Ciencias de la Computación, ETSI Informática. Universidad de Málaga, Bulevar Louis Pasteur, 35. Campus de Teatinos. 29071 Málaga. España.
Omar Cuevas Salazar, Ph.D., Professor, Instituto Tecnológico de Sonora, campus Náinari, Antonio Caso 2266, Villa Itson, 85130, Ciudad Obregón, Sonora, México.
Daniel Mocencahua Mora, Ph.D., Professor, Benemérita Universidad Autónoma de Puebla, 4 sur 104 Centro Histórico, 72000, Puebla, Puebla, México.
Maria Esther Velarde Flores, Ph.D., Professor, Instituto Tecnológico de Sonora, Campus Náinari, Antonio Caso 2266, Villa Itson, 85130, Ciudad Obregón, Sonora, México.

theory support the interpretations of test results for the purpose that was developed for. Validity, has five different evidence sources: (a) evaluation process; (b) test content; (c) response process; (d) internal structure; (e) relationship with other variables and consequences for the subject of evaluation. Within the content validity, an important aspect is the exam questions' quality (Rivera, Flores, Alpuche, & Martínez, 2017).

Validation also includes estimates of the characteristics of analytical performance and diagnostic testing. However, for a test to remain authentic a careful monitoring of its performance is needed, often by regulating internal controls behavior through a period of time. This ensures that the test, according to the original validation, always maintains its performance characteristics (Márquez, Ramos, & Lopez, 2015). Nowadays, content validity or item validity is considered a necessary (not enough) condition to make test scores interpretations. In addition, content validity not only refers to measuring instrument items but to its instructions and score criteria (Pedroza, Suárez, & García, 2013).

The purpose of this study is to perform the validation of mathematics test items' to be applied during the admission process of higher education new students. According to Niño, Hakspiel, Mantilla, Cardenas, and Guerrero (2017), counting with this information about students' math abilities in the first stages of their formative experience, can save highly cost mistakes, contribute to math abilities assesment field development and the training of specialists in this área. In spite of the implementation in several countries of this kind of validation since last century early years, there is still a gap with other countries that are not experimenting the benefits of its aplication.

Theoretical Framework

The basic measurement theory had place in the eighteenth century and since the nineteenth century these ideas were applied in the educational field in countries like Germany, England, and the United States of America (Rizo, 2001). During the second half of the twentieth century, the item response theory (TRI) was conceived by Psychology gurus, with the spread of computers statistical models like Birnbaum's (1957) and Rasch (1960) were developed. Lord and Novick launched, in 1968, Statistical theories of mental test scores. The TRI tries to give a probabilistic basis of the problem of measuring features and unobservable constructs (latent traits), considering the item, not the overall score, as the basic unit of analysis (Macías, 2011).

Now the measuring range of the instrument is set by TRI, based on Attorresi, Lozzia, Abal, Galibert, and Aguerri (2009) its substantial objective is the construction of measuring instruments with invariant properties between populations. If two individuals have the same level of measured feature they will have equal probability of giving the same response, regardless of the population of belonging.

The established scale goes from -3 to +3, according to the characteristic of the item (CCI) whose behavior tends to $-\infty$ to $+\infty$. However, the ICC depends on three basic parameters, or parameters of the item. These are the difficulty index, the discrimination index, and the random index (or pseudo-random), so the maximum and minimum is taken as a reference.

Rojas, Manriquez, and Gatica (2004) affirm that a TRI model assumes that there is a latent variable (θ), that can not be directly observed, and that must be estimated for each subject from his/her answers provided in the measuring instrument. In addition, the scale must be aligned to the TRI, where the range chosen corresponds to the θ value in the first question response (see Table 1).

The scale established is used to determine the input knowledge level of the student. He/she chooses, with the subjective perception of his/her own ability, an initial starting level from a set of qualitative values preset by

expert analysis and parameters marked by TRI. Also, Debera, and Nalbarte (2006) established a scale of -2 to +2, this after an analysis of the results is shown in the pretest and adjusting skill levels evaluated for application and adaptation in the final test.

Table 1

Student Skill Level Scale

Student skill level		
Num.	Skill description	Ability
1	I know nothing about the topic	-3
2	I only know something about the topic	-2
3	I read earlier about the topic	-1
4	I have studied the topic before	0
5	I've seen the topic before, but I need to refresh my knowledge	1
6	I have dominated the topic	2
7	I am an expert on the topic	3

Another case is studied by Guzman (2005), which has four experiments with different amounts of item banks, those having more than 400 items used a scale of -2 to +2, in the case of those that have 300 items level of knowledge ranged between -4 and 4 and were generated following a normal distribution. The response model used is the 3pl consisting in three parameters that define the characteristics of each item (a) item discrimination (a parameter, is a parameter that measures the ability of the item to distinguish subjects depending on its level at latent trait), (b) the item difficulty (parameter b), and (c) pseudo-guessing (parameter c indicates the possibility that a subject can hit the item by luck).

On the other hand, Hidalgo-Montesinos and French (2016) state that the analysis of the items through the TRI is a system validation testing and for this there is software to perform psychometric analyses based on this theory. Sympson and Haladyna (1988) develop multiple weighting methods for tests related to a domain. This procedure ponders each item according to the average percentile of examinees who chose that option. The results show that this method multiple weighting yields the highest test reliability and the best domain related validity.

Meanwhile, Razel and Eylon (1987) validate different ways to qualify Raven's Colored Progressive Matrices Test. The autors compare the conventional way of describing the evidence against three scoring methods of multiple weighting: (1) according to the theory of cognitive processing, (2) according to expert opinion, and (3) based on the responses selected by students.

Rasch model, proposed in 1960 is based on the assumptions that the attribute to be measured can be represented in a single dimension which would place people and items together. In addition, the level of the individual in the attribute and item difficulty determine the probability that the answer is correct. If situation control is right, this expectation is reasonable and well chosen to represent the mathematical model. Rasch used the logistic function to model the relationship. Equation indicates that the ratio of the probability of a correct response and the probability of an incorrect response to an item, is a function of the difference in the attribute between the level of the person (θ_s) and the item level (β_i). So, $\ln\left(\frac{P_{is}}{1-P_{is}}\right) = (\theta_s - \beta_i)(\frac{P_{is}}{1-P_{is}})$

Cronbach's Alpha calculated in Rasch model, is an indicator of reliability in terms of internal consistency for an instrument. It is an indicator with which the accuracy of the test in terms of its internal consistency is measured and points to the score's stability level. It is estimated what proportion of the observed variability in

the scores corresponds to true variance, variance due to differences in the construct to be measured. Its maximum value is 1, the closer it gets to this higher value is the level of reliability (Jimenez & Montero, 2013).

Overall, international programs of educational tests deemed acceptable Alpha values when the number is bigger than 0.8, but authors like Nunnally and Bernstein (1995) are stricter when it comes to high-stakes testing for taking direct decisions and indicate that Alpha must be at least 0.9. On the other hand, if it comes to instruments that will be used only for research processes criteria can be flexible. In that case acceptable Alpha values would be equal to or bigger than 0.7 (Jimenez & Montero, 2013).

The results show that the weighted rating is preferable to conventional (0 to 1) because it improves the validity and reliability of the test: empirical weighting is the best method. In addition, the weighting and validation of Multiple Choice Questions (POM), responses are optimal because all options are partially correct, but only one fits precisely (Jara, 2015).

Moodle is an open and modular software code that, in addition to content management tools, and offers several methods of authentication between users. The data handled are stored on the platform server in a database, and are organized into different structures. Moodle main structures are: users, courses, and modules. Modules can be classified into three groups: resources, activities, and blocks (Presedo, Armendariz, López-Cuadrado, & Perez, 2015). In the second, questionnaires, tests, or examinations are established. This segment is the results of the items validation used in this investigation.

Method

The study involved 940 new students in the area of engineering of Instituto Tecnológico de Sonora. The instrument is a diagnostic mathematics test that consists of 22 items divided into seven topics: algebraic expressions grade 1, grade 2 and higher level, rational expressions, circumference, basic trigonometry, and exponential and logarithmic equations. Each item is multiple choice, consisting of a statement and four options with only one correct answer.

Phase 1: Validation by experts

A work plan was established with an expert in the area of mathematics from Instituto Tecnológico de Sonora in collaboration with an investigator in the study of Computerized Adaptive Test development from the University of Málaga. The validation by experts was conducted through an observation guide and the data obtained by the instrument were analyzed by matching attributes aligned to the research objectives. With the results of the observations, required changes to items were done. The observation guide described research objectives and was based on a Likert ranging, from 1 (very few) to 5 (very acceptable), the criteria to be evaluated were writing clearness, clear layout, contribution with assessment objectives, theoretical and empirical correspondence, completeness of the content, internal consistency among the items, evaluation of significant learning contribution to measure established constructs, and a space for observations was provided. The results showed that all criteria had a very acceptable rating.

Phase 2. Pilot testing

The test was applied through Moodle with a limit of 1 hour to be answered. Participants were divided into groups comprising of 90 students approximately, due to the number of computers available. Therefore, the application contemplated 10 groups that anwered the test during one day from 7 am to 6 pm.

Phase 3. Validity and reliability from the results

After running the pilot testing, the difficulty levels of the items, discrimination, biserial correlation, and reliability were analyzed through the development of a Rasch model supported in R and SPSS software.

About the items evaluation, based on Feedback (2018), the difficulty level of an item was analyzed, which is determined by the proportion of students with correct answers to each question in relation to the total group. This index shows how easy (values close to 1) or difficult (values close to 0) the question has been for the total population.

Mathematically: $\text{Difficulty level} = \frac{\text{number of correct answers}}{\text{total number of students}}$

In addition, the discrimination level was obtained by biserial correlation (r_{bis}), which determined the level in which competencies measured by the test are also being measured by the item. The r_{bis} provides an estimate about the Pearson product-moment correlation between the total score of the test and the item hipotetical continuos, when this is dichotomized into right and wrong answers (Henrysson, 1971).

However, there is also the point biserial correlation (r_{pbis}), which established the adequacy of people with correct answers, how much predictive power the item has, and how it can contribute to predictions. Based on Henrysson (1971) the r_{pbis} indicates more about predictive validity of the test than the biserial correlation coefficient, since it tends to favor medium difficulty items. It is also suggested that the r_{pbis} is a measure that combines the relationship between the item criteria and difficulty level. The test depends on its own difficulty for being adaptative, so the calculus r_{pbis} was chosen, as is presented in the next equation:

$$r_{pbis} = \frac{\bar{x}_1 - \bar{x}_0}{S_x} * \sqrt{\frac{n_1 n_0}{n(n-1)}}$$

r_{pbis} = biserial correlation point

$\bar{x}_1$ = Average of total scores of those who correctly answered the item.

$\bar{x}_0$ = Average of total scores of those who answered the item incorrectly.

S_x = Standard deviation of the total scores.

n_1 = Number of cases with correct answer to the item.

n_0 = Number of cases with incorrect answer to the item.

$n = n_0 + n_1$

Results

An analysis of difficulty level for each question and for the test in general was made, the results are presented in Table 2.

Table 2

Difficulty Level per Question

Question	Correct answer	Difficulty level	Question	Correct answer	Difficulty level
1	917	0.97	12	667	0.70
2	715	0.76	13	556	0.59
3	541	0.57	14	649	0.69
4	434	0.46	15	480	0.51
5	544	0.57	16	414	0.44
6	507	0.53	17	237	0.25
7	255	0.27	18	854	0.9
8	774	0.82	19	548	0.58
9	807	0.85	20	566	0.6
10	592	0.62	21	359	0.38
11	523	0.55	22	538	0.57

The results indicate that the higher difficulty level (0.25) was found in the topic about algebraic expressions Grades 4 and 5, followed by 0.27 for the absolute value exercise. In contrast, the starting question about point location has the lowest difficulty level with 0.97. In relation to the general difficulty level of the test, a 0.59 grade was obtained. The test's general average was 6.03. Figure 1 shows the frequencies of each level of difficulty and their normal behavior, this suggests that the test is balanced.

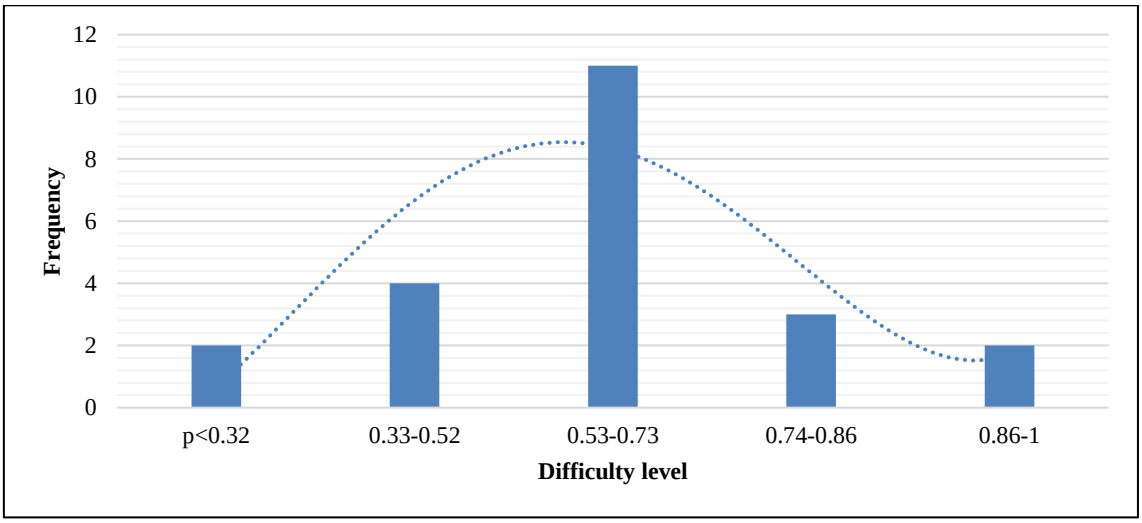

Figure 1. Representation of test's difficulty levels frequencies.

Data above indicate that the test is appropriate, according to Backhoff, Larrazolo, and Roses (2000), the half level of test difficulty must range between 0.5 and 0.6; *p* values are distributed as follows: 5% of difficulty, 20% fairly easy, 50% with average difficulty, 20% moderately difficult, and 5% difficult.

From calculating the biserial correlation point using Excel software, the following results were obtained (see Table 3).

Table 3

Index Discrimination and r_{pbis}

Question	High (only 27%)	Low (only 27%)	n_1	n_0	r_{pbis}	Question	High (only 27%)	Low (only 27%)	n_1	n_0	r_{pbis}
1	82	165	917	37	0.21	12	76	104	667	287	0.49
2	78	115	715	239	0.46	13	71	79	556	398	0.53
3	66	79	541	413	0.53	14	76	99	649	305	0.5
4	53	64	434	520	0.53	15	63	66	480	474	0.54
5	71	76	544	410	0.53	16	58	54	414	540	0.53
6	59	78	507	447	0.53	17	33	31	237	717	0.46
7	35	3. 4	255	699	0.47	18	82	149	854	100	0.33
8	79	130	774	180	0.42	19	73	75	548	406	0.53
9	81	137	807	147	0.39	20	72	80	566	388	0.53
10	70	89	592	362	0.52	21	54	43	359	595	0.52
11	71	70	523	431	0.53	22	74	71	538	416	0.53

To obtain the biseral correlating point, the standard deviation is 0.42, average of correct answer 0.45, average of wrong answer 0, results according to each item's level from 0.45 (correct response) to 0 (incorrect answer was used); therefore, the calculation of r_{pbis} has an average of 0.48. As Backhoff et al. (2000) say, if the r_{pbis} is less

than 0 it is negatively discriminated poorly from 0 to 0.14, from 0.15 to 0.25 is regular, is considered a good discriminative power from 0.26 to 0.35, and 0.35 and above is an excellent level. So, unlike the discrimination test only one question is at the regular level. Similarly, the data produced by the same platform used for the application of the test, established that the rate of feasibility has an average of 0.59, the discrimination index 0.39, and 0.52 for the discriminatory efficiency. So that, it can be inferred that the test is suitable for the application. In the case of reliability, the Rasch model in SPSS software generated the data shown in Table 4.

Table 4

Descriptive Statistics and Reliability Through Rasch Model in SPSS

Number of cases	Number of items	People's reliability index	Item's reliability index
940	22	0.958	0.754
	Scores	Estimated measure (people)	Standard error of estimation
Mean	12.06	-0.03	0.31
Standard deviation	3.74	0.446	0.042

The reliability index of the items is between 0.734 and 0.775 with the base of the cases. Once detecting items and people who did not meet the expectations of the model, they were removed from the original database and the analysis was ran again. New reliability statistics showed a value of 0.74, indicating that the estimatimations of Rasch difficulty parameter were consistent. The people reliability index increased to 98.77%, in the case of estimation error it decreased to 1.2%. The characteristic curves of the items (CCR) were established as shown below, relating people's measure and their chance of answering the item correctly.

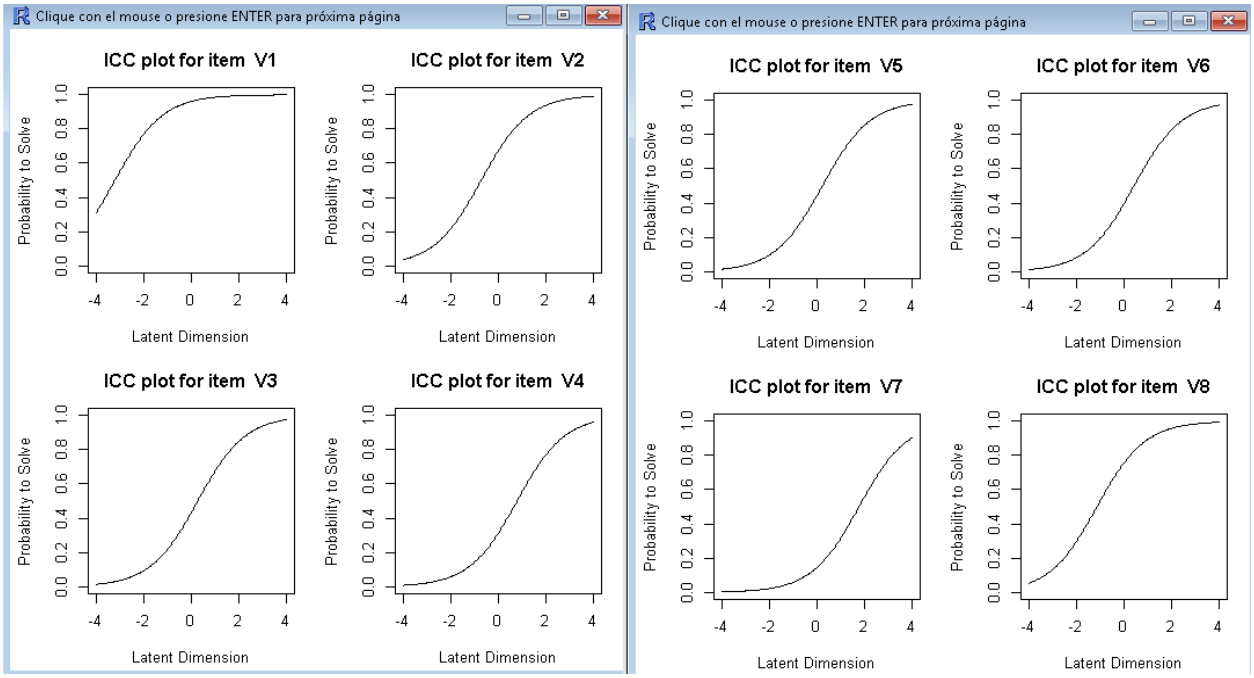

Figure 2. CCR of items 1, 2, 3, 4, 5, 6, 7, and 8.

The CCR of item 1 demonstrates its behavior perfectly, because it is the easiest item, the possibility to be answered in a difficulty level of +4 is 1 and therefore is higher than -4 to 0.3. In exercises 2, 3, 4, 5, and 6, the CCR is presented uniformly, so then, to respond with a value -4 is the same for all three items and the same case

is presented in the difficulty of +4 whose probability is 1. Item 7 has a behavior that tends to the highest difficulty, this can be proved with the calculated index in SPSS (0.27). Item 8 behaves with probability of low difficulty, so that, the data tend to locate in the positive side.

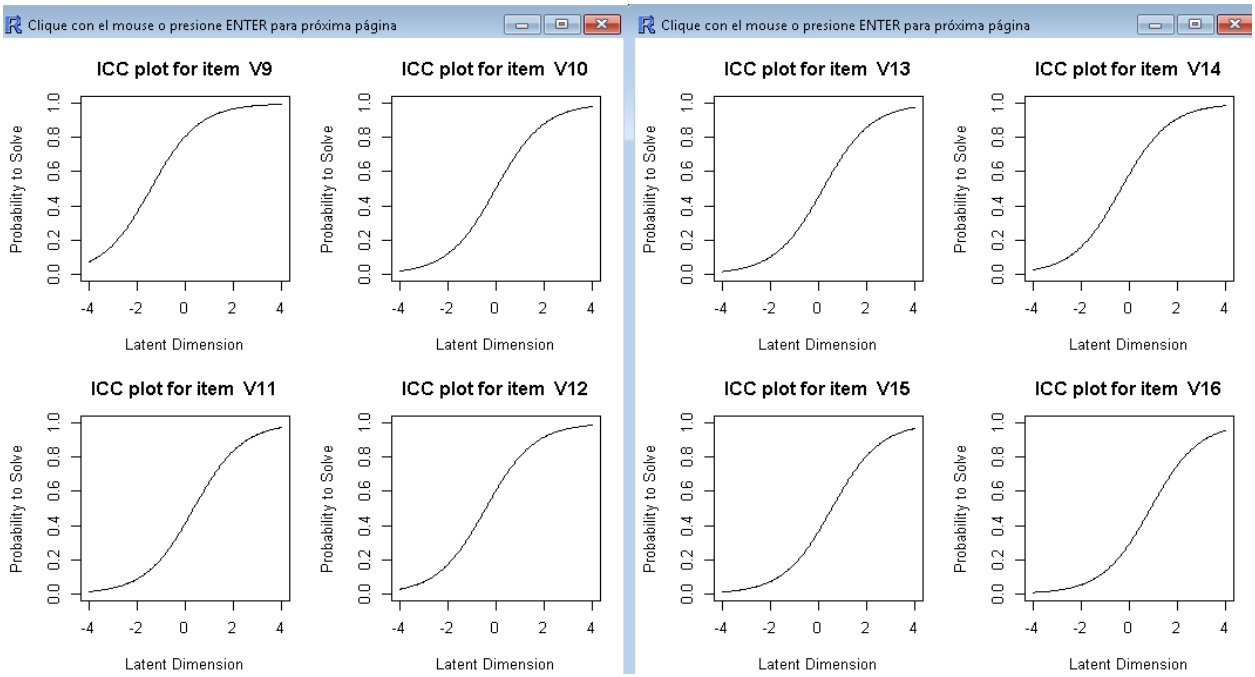

Figure 3. CCR of items 9, 10, 11, 12, 13, 14, 15, and 16.

For items 9 and 12 there is a higher chance to be answered correctly, item 16 increases the probability of an incorrect answer. For the remaining items probability is average, students who solve exercises have the same chance of answering correctly or incorrectly.

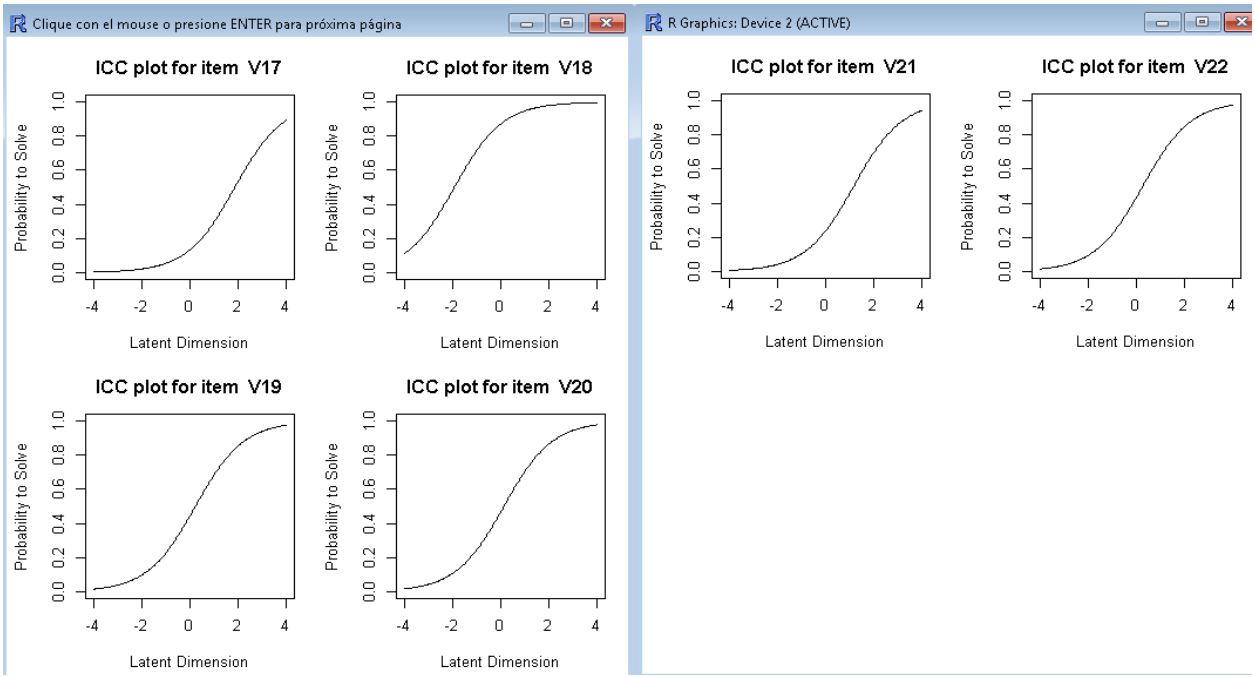

Figure 4. CCR of items 17, 18, 19, 20, 21, and 22.

In Figure 4, CCR of items 17 and 21, represents a bigger chance for student to respond in direction to negative values of difficulty, item 18 established that the probability of a correct answer is higher. The remaining items have an average probability. With Rasch model it can be confirmed that easiest items are 1 and 8 and the hardest ones are 7 and 17 with an index below from 0.25. The mayor part of the items is located in a difficulty index from 0.5 to 0.7, so then the test is suitable for its application.

Conclusions and Recommendations

Rasch model application represents to obtain the reliability level of those being tested, of the test itself and the items consistency. There is a precise identification of the items' behavior, allowing establishing if a restructure of items is needed or if the test is ready for its application to the population that is being studied.

In the CCR of the items it can be seen that the skill level average of the examinees is very nivelated to the item difficulty average, most of the population was located above, this indicates that the test was slightly difficult for examinees. But the majority of the items are concentrated on intermediate skill levels, there are almost none items that provide information at high levels or low levels (only two in each cases). A startup test or diagnostic evaluation is desirable to provide items with optimal levels in technical quality in all skill levels and with the attributes described in the experts evaluation.

Furthermore, the validation through Rasch assumes that the test is always measured with the same reliability, it is known that the parameters of the subjects with low ability levels would be estimated more accurately, or in the case of examinees located in high skill levels, with them the parameters of the difficult items could be estimated more accurately. Because of that, it is necessary to identify the difficulty and discrimination index with some other methods that support what is visualized in the model. In this analysis it was approached through the methodologies proposed by Backhoff et al. (2000), identifying that the items are appropiate to the test and only one of them should have an adjustment in its structure.

The difficulty level of the test was 0.59, so a positive asymmetry is established congruent with what is wanted in a test. It was observed that there are items in all the range of difficulty and not only those focused on 50% of difficulty, in order to measure the ability level of each person accurately and with equal opportunities. According to Hurtado (2018) medium difficulty tests show better results because they represent more accurate scales.

In the correlation biserial point, average is located at 0.48, so based on theory the test has an excellent level. From the data obtained by Moodle platform, it shows an average of feasibility index in 0.59, discrimination index of 0.39, and discriminatory efficiency of 0.52, and it can be established that the test is suitable for its application.

It is necessary to continue studying in detail the processes involved in items solution from the list of each item responses (in case they answered the item), being able to identify which processes represent different levels of difficulty in the items and to glimpse the characteristics shared by items of a similar difficulty level, generating item construct indicators and processes that must be done to get difficulty levels pursued and develop more adapted tests for the particular needs.

References

Attorresi, H., Lozzia, G., Abal, F., Galibert, M., & Aguerri, M. (2009). Teoría de Respuesta al Item. Conceptos básicos y aplicaciones para la medición de constructos psicológicos. *Revista Argentina de Clínica Psicológica, XVIII*(2), 179-188. Retrieved from http://www.redalyc.org/html/2819/281921792007/

Backhoff, E., Aguilar, J., & Larrazolo, N. (2006). Metodología para la validación de contenidos de exámenes normativos. *Revista Mexicana de Psicología, 23*(1), 79-86. Retrieved from www.redalyc.org/pdf/2430/243020646010.pdf

Backhoff, E., Larrazolo, N., & Rosas, M. (2000). Nivel de dificultad y poder de discriminación del Examen de Habilidades y Conocimientos Básicos (EXHCOBA). *Revista Electrónica de Investigación Educativa, 2*(1), 10-29. Retrieved from http://redie.uabc.mx/vol2no1/contenido-backhoff.html

Birnbaum, E. (1957). The cost of a foreign exchange standard or of the use of a foreign currency as the circulating medium. *IMF Staff Papers, 5*(3), 477-491.

Cronbach, L. J. (1971). Test validation. In R. Thorndike (Ed.), *Educational measurement* (2nd ed., p. 443). Washington DC: American Council on Education.

Debera, L., & Nalbarte, L. (2006). Pruebas diagnósticas: una aplicación a la teoría de respuesta al item, aproximación clásica y bayesiana. Instituto de Estadística. F.C.E. y A. Universidad de la República. Retrieved from http://www.iesta.edu.uy/wp-content/uploads/2010/03/0601.pdf

Ebel, R. L., & Frisbie, D. A. (1986). *Essentials of education measurement*. Englewood Cliffs, NJ: Prentice Hall.

Feedback. (2018). Conceptos básicos sobre Análisis de Item, Datos Estadísticos y Calificación Sustractiva. Retrieved from http://www.sistemafeedback.com.ar/descargas/Evaluando_al_Examen.pdf

Godino, J., Batanero, C., & Font, V. (2003). *Fundamentos de la enseñanza y el aprendizaje de las Matemáticas para maestros. Matemáticas y su didáctica para maestros* (p. 79, 81). España: Departamento de Didáctica de la Matemática, Facultad de Ciencias de la Educación, Universidad de Granada. Retrieved from https://www.ugr.es/~jgodino/edumat-maestros/manual/1_Fundamentos.pdf

Guzmán, D. (2005). Un modelo de evaluación cognitiva basado en Tests Adaptativos para el diagnóstico en Sistemas Tutores Inteligentes (PhD thesis, Recuperada de Universidad de Málaga, 2005). Retrieved from http://www.lcc.uma.es/repository/fileDownloader?rfname=LCC1406.pdf

Henrysson, S. (1971). Gathering, analysing, and using data on test items. In R. L. Thorndike (Ed.), *Educational measurement*. Washington, DC: American Council on Education.

Hidalgo-Montesinos, M., & French, B. (2016). Una introducción didáctica a la Teoría de Respuesta al Item para comprender la construcción de escalas. *Revista de Psicología Clínica con Niños y Adolescentes, 3*(2), 13-21. Retrieved from http://www.revistapcna.com/sites/default/files/16-11.pdf

Hurtado, L. (2018). Relación entre los índices de dificultad y discriminación. *Revista Digital de Investigación en Docencia Universitaria, 12*(1), 273-300. Retrieved from http://www.scielo.org.pe/scielo.php?script=sci_arttext&pid=S2223-25162018000100016

Jara, M. (2015). Validez y confiabilidad en la construcción de reactivos utilizados en pruebas de opción múltiple (POM). Retrieved from https://www.researchgate.net/publication/282367035

Jiménez, K., & Montero, E. (2013). Aplicación del modelo de Rasch, en el análisis psicométrico de una prueba de diagnóstico en matemática. *Revista digital Matemática, Educación e Internet, 13*(1), 1-24. Retrieved from https://tecdigital.tec.ac.cr/revistamatematica/ARTICULOS_V13_N1_2012/RevistaDigital_Montero_V13_n1_2012/RevistaDigital_Montero_V13_n1_2012.pdf

Leyva, Y. (2011). Una reseña sobre la validez de constructo de pruebas referidas a criterio. *Perfiles Educativos, 33*(131), 131-154. Retrieved from http://www.scielo.org.mx/scielo.php?script=sci_arttext&pid=S0185-26982011000100009

Lira, R. I. (2013). Diseño de un enfoque modélico de evaluación institucional para el Instituto Tecnológico de Costa Rica: Construcción y validación de los indicadores de proceso y de Producto de las áreas de formación universitaria y de impacto que forman dicho modelo (PhD thesis, Universidad de Valencia, España, 2013). Retrieved from http://roderic.uv.es/handle/10550/27379

Macías, E. (2011). Validación y confiabilidad de pruebas de opción múltiple para la evaluación de habilidades (M.Sc. thesis, Centro de investigación en Matemáticas, 2011). Retrieved from https://cimat.repositorioinstitucional.mx/jspui/bitstream/1008/245/2/TE%20373.pdf

Márquez, A., Ramos, Á., & López, A. (2015). Qué sabe Ud. acerca de… la validación de pruebas diagnósticas? *Revista Mexicana de Ciencias Farmacéuticas, 46*(3), 86-90. Retrieved from http://www.redalyc.org/pdf/579/57945705010.pdf

Niño, L., Hakspiel, M., Mantilla, L., Cárdenas, M., & Guerrero, N. (2017). Adaptación y validación de instrumento para evaluar habilidades psicosociales y hábitos saludables en escolares. *Universidad y Salud, 19*(3), 366-377. Retrieved from http://www.scielo.org.co/pdf/reus/v19n3/0124-7107-reus-19-03-00366.pdf

Nunnally, J. C., & Bernstein, I. H. (1995). Chapter 3: Validity. *Psychometric Theory, 3*(1), 83-103.

Pedrosa, I., Suárez, J., & García, E. (2013). Evidencias sobre la validez de contenido: avances teóricos y métodos para su estimación. *Acción psicológica, 10*(2), 3-18. Retrieved from http://scielo.isciii.es/scielo.php?script=sci_arttext&pid=S1578-908X2013000200002

Presedo, C., Armendariz, A., López-Cuadrado, J., & Pérez, T. (2015). Calibración de items vía expertos utilizando Moodle. *Revista Iberoamericana de Educación, 69*(1), 117-132. Retrieved from https://rieoei.org/RIE/article/view/158

Prieto, G., & Delgado, A. (2003). Análisis de un test mediante el modelo de Rasch. *Psicothema, 15*(1), 94-100. Retrieved from http://www.psicothema.com/pdf/1029.pdf

Razel, M., & Eylon, B. S. (1987). *Validating alternative modes of scoring for coloured progressive matrices.* American Educational Research Association, Washington.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests.* Copenhagen: Denmarks Paedagogiske Institut.

Rivera, J., Flores, F., Alpuche, A., & Martínez, A. (2017). Evaluación de reactivos de opción múltiple en medicina. Evidencia de validez de un instrumento. *Investigación en educación médica, 6*(21), 8-15. Retrieved from http://www.scielo.org.mx/scielo.php?script=sci_arttext&pid=S2007-50572017000100008&lng=es&nrm=iso

Rizo, F. (2001). Evaluación educativa y pruebas estandarizadas. Elementos para enriquecer el debate. *Revista de la Educación Superior, 30*(120), 1-12. Retrieved from http://publicaciones.anuies.mx/pdfs/revista/Revista120_S3A3ES.pdf

Rojas, M., Manríquez, G., & Gatica, Y. (2004). Curso de UML Multiplataforma Adaptativo Basado en la Teoría de Respuesta al Item. *Revista Ingeniería Informática, 10*, 1-10. Retrieved from http://inf.udec.cl/~revista/ediciones/edicion10/psalcedo01.pdf