Proceedings CIEAEM 61 – Montréal, Quebéc, Canada, July 26-31, 2009
"Quaderni di Ricerca in Didattica (Matematica)", Supplemento n. 2, 2009.
G.R.I.M. (Department of Mathematics, University of Palermo, Italy)

Process of design and validation of the Mathematical Essential Aptitudes Diagnosis Test for first-year university students: results from a pilot run of the test

Evelyn Agüero-Calvo[1]
Technological Institute of Costa Rica – University of Valencia

Abstract

At a higher education institute in Costa Rica was performed a research work which intends to begin a process of measurement of the mental abilities that students possess when they enter the university. The study was focused on the design and validation of a repeated-use test, using processes of items analysis, expert's judgements and pilot runs of the test. The promotion of projects like this is important in order to improve the methods of design and development of tests by mathematics teachers and to promote high quality pertinent activities.

Main idea

This article documents the first experience from a research work which focused on the pilot design and validation of a repeated-use test that will diagnose the possession of mathematical aptitudes in first-year university students in order to predict academic success and guide educative interventions.

The referred research intends to begin a process of measurement of the mental abilities that students possess when they enter the university. Those abilities are supposed to have been developed and reinforced through the elementary and secondary school.

But in Costa Rica, as in many other countries, there is a big problem with the mathematical education because the students do not have very good mathematical thinking and reasoning, which harms their social, labour and personal development. A mathematics professor can notice the lack of connected representations and necessary skills among university students who take examinations semester after semester, course after course.

At a higher education institute like the Technological Institute of Costa Rica, it must become a priority to establish the profiles and academic possibilities that the students who have been admitted have, in order to give them follow-up activities and support their academic success.

Establishing a diagnosis process before the students fail in their courses, at the moment when they enter the university and begin with their first courses, might mark a signifi-

[1] Mathematics professor at the Technological Institute of Costa Rica. Doctorate student at the University of Valencia.

cant difference if the results of the diagnosis test are used to organize the educational process by improving, planning, programming and making decisions (Escudero, 2003).

Methodology

The study began with the design of 54 items or questions representing mathematical abilities such as identification, mental representation, codification, classification, analysis, synthesis, hypothetical reasoning, logical inference, analogical reasoning and syllogistic reasoning for example (MEP, 2001). These 54 items shaped the bank of items and were checked in a process of expert's judgement.

The revision of the items was performed by a group of nine mathematics teachers, called experts, who evaluated each item in five categories: clarity, bias, representation, discriminatory capacity, and adequate distracters. These qualifications were tabulated and analyzed with the statistical package SPSS 13.0, calculating the Kendall's *W* (also known as Kendall's coefficient of concordance) which is a non-parametric statistic that is used for assessing agreement among raters. Kendall's *W* ranges from 0 (no agreement) to 1 (complete agreement). Intermediate values of *W* indicate a greater or lesser degree of unanimity among the various responses.

For example, the clarity of the items that represent the mental ability of identification was evaluated by the experts. The results of the computations for Kendall's *W* are:

Ranks

	Mean Rank
Clarity of F1R1	2.17
Clarity of F1R2	1.50
Clarity of F1R3	**2.33**

Test Statistics

N	9
Kendall's *W*	**0.467**
Chi-Square	8.400
df	2
Asymp. Sig.	0.015

The value of 0.467 represents a moderate concordance between the raters; then the third item (F1R3) is clearer than the other two. In the same form, these three items (F1R1, F1R2, F1R3) were analyzed in the remaining four categories. At the end, the third item (F1R3) was one of 18 items chosen to shape the pilot test that was applied to a first-year university students' sample.

However, the selection of all the items was not as easy as in the example shown. There were 18 items in which there no was an acceptable value of concordance for Kendall's *W*. These items were modified or replaced, for which a new expert's judgement run was necessary in order to obtain a better value of concordance.

With all the items selected to shape the pilot test, a pilot run was performed with 147 first-year university students at the Technological Institute of Costa Rica. Their answers were codified in order to make a statistical analysis of the test and the items.

Some examples of the tasks included in the test are the following:

Which pair of numbers completes the next numerical sequence?

1 10 3 9 5 8 7 7 9 6 ? ?

a) 11 5
b) 10 5
c) 10 4
d) 11 6

For each function $f(x) = x^n$, with n a real number and $n \neq -1$, the function F is defined as $F(x) = \frac{x^{n+1}}{n+1}$; so, for the function $f(x) = \frac{1}{x^9}$ it is true that:

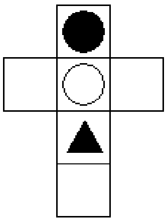
a) $F(x) = \frac{10}{x^{10}}$

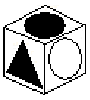
b) $F(x) = \frac{x^{10}}{10}$

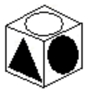
c) $F(x) = \frac{-8}{x^8}$

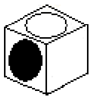
d) $F(x) = \frac{-x^{-8}}{8}$

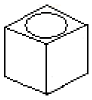
Which of the boxes below can be formed by folding the figure on the right?



a) 

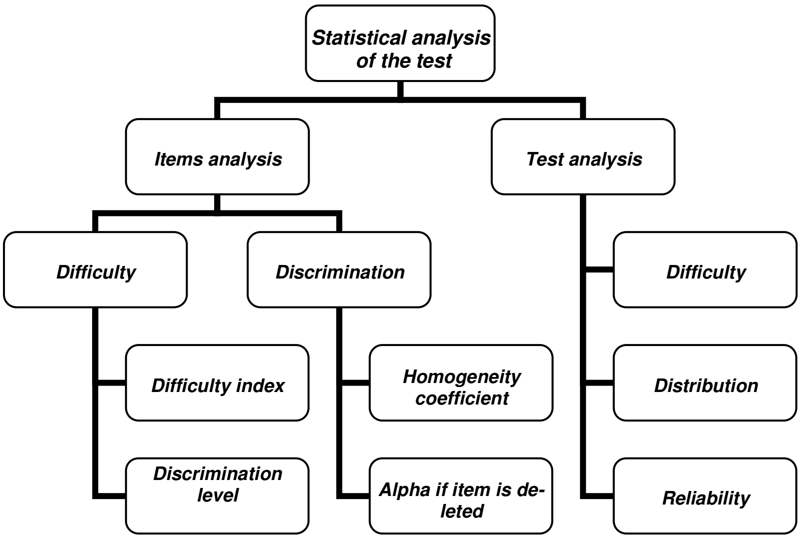
b) 

c) 

d) 

The statistical analysis of the items and the test is necessary in order to see if the test has internal consistency and reliability. The results of these analyses are determinant to decide if the test measures what it must measure according to the objectives of its design. If the results are not satisfying, the test has to be redesigned and the necessary adjustments made.

The statistical analysis of the test was performed following the next diagram:



Discussion of results

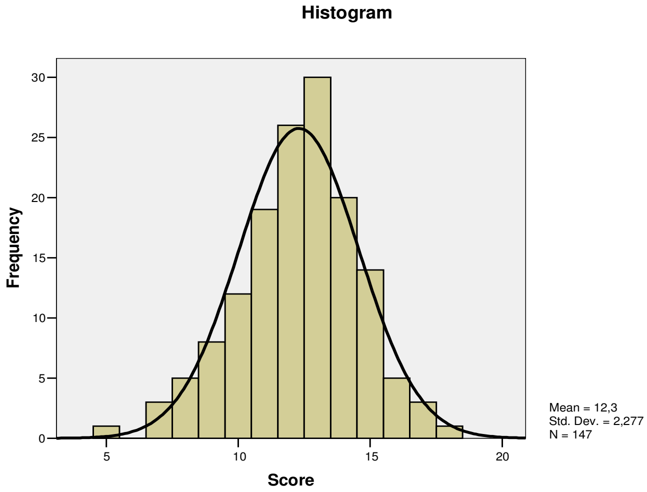
Each indicator of the above diagram was calculated with SPPS 13.0, and a summary of the outcomes is the following:
- Difficulty index: for this indicator we use the mean of each item. The results were 50% of easy items, 44.4% of middle items and 5.6% of difficult items. So, more difficult items have to be included in next versions of the test.
- Discrimination level: for this indicator we use the variance of each item. The results were 50% of items with critical discrimination, 17% items with null discrimination and 33% items with optimum discrimination. So, more items with optimum discrimination are needed.
- Homogeneity coefficient: for this indicator we use the corrected item-total correlation. The results were 28% items with middle correlation with the total score of the test, and 72% items with low correlation. So, the majority of the items need a revision.

- Alpha if item is deleted: for this indicator we use the Cronbach´s alpha value if the item is deleted. The results were that 15 items favor the reliability of the test and three items damage the reliability of the test. So, these three items have to be eliminated.

- Test difficulty: for this indicator we use the theoretical mean of the test which, in this case, is 9 because there are 18 questions or items, and the real mean obtained which is 12.3.

Thus, the test is easy because the real mean is greater than the theoretical mean, so the test has to be more difficult.

- Test distribution: for this indicator we use the values of skewness and kurtosis, which are respectively -0.337 and 0.291. So, the distribution of the score of the test is negative and mesocortical. Also, by viewing the histogram of the score of the test, it is noted that the test distribution fits the normal distribution, which indicates that the test generates good variability.



- Test reliability: for this indicator we use the Cronbach´s alpha value; in this case it is 0.483 which represents a middle reliability. A middle reliability is not enough for a cognitive test like this, which intends to measure mathematical abilities. So, the reliability of the test has to be increased by eliminating the items that damage the reliability.

Conclusions

In order to continue with the design of the test proposed in this work and according to the statistical analysis performed, many adjustments need to be made. More items in the bank of items, more experts to revise the items, pilot test with more items or questions, more institutions to perform runs of the pilot test, and more students in the samples, are some of the suggestions to get better results.

Designing a valid and reliable test is not an easy job, and it is even more difficult if the test is wanted for specific purposes and, as in this case, for measuring mathematical abilities. That's the reason why the processes of analyzing the items in categories through an expert's judgement and making pilot runs of the test are important for the research (Ruiz-Primo, Jornet & Backhoff, 2006).

The outcomes of this research can help to improve the methods of design and development of tests by mathematics teachers and to promote high quality pertinent activities in order to help students reach the maximum of their capacities (Primi, Angeli & Medeiros, 2002). It is important to raise awareness among academic decision makers and government authorities about the problem in mathematical education (Davis & Barnard, 2000).
Also, with the results of research projects like this, reliable indicators and instruments can be generated to help other researchers and orientate any kind of actions and ameliorate programs in order to improve the quality of mathematical education and, not far from here, create a system of evaluation of Costa Rican education.

Related references

Davis, E. & Barnard J. T. (2000). What seems to be happening in mathematics lessons? Findings from one school system and five student teachers. *The Mathematics Educator*, 10 (1), 11-18. Consulted on December 12, 2007 on http://math.coe.Uga. edu/tme/v10n1/3davis.pdf
Escudero, T. (2003). Desde los tests hasta la investigación evaluativa actual. Un siglo, el XX, de intenso desarrollo de la evaluación en educación. *Revista Electrónica de Investigación y Evaluación Educativa*, 9 (1), 11-43. Consulted on December 12, 2007 on http://www.uv.es/RELIEVE/v9n1/RELIEVEv9n1_1.htm
Ministerio de Educación Pública (MEP). (2001). *Programa de Estudios en Matemática de la Educación Diversificada*. San José, Costa Rica: Publicaciones del Ministerio de Educación Pública.
Primi, R., Angeli, A. & Medeiros, C. (2002). Habilidades básicas e desempenho acadêmico em universitários ingressantes. *Estudos de Psicologia*, 7 (1), 47-55. Consulted on December 14, 2007 on http://www.scielo.br/pdf/epsic/v7n1/10953.pdf
Ruiz-Primo, M. A., Jornet, J. M. & Backhoff, E. (2006). *Acerca de la validez de los exámenes de la calidad y el logro educativos*. México, DF: Instituto Nacional para la Evaluación de la Educación. Consulted on June 1, 2007 on http://www.inee.edu.mx/images/stories/documentos_pdf/Publicaciones/Cuadernos_tecnicos/ct20_sobre_validez_excale.pdf