Revista
Torreón Universitario
www.faremcarazo.unan.edu.ni / ISSN 2410-5708 / e-ISSN 2313-7215
Year 7 | No. 18 | p. 93 - p. 105 | february - may 2018

Explanations for the presence of atypical patterns of response of a knowledge evaluation test in the FAREM-Carazo

PhD. Eduardo Doval
Universidad Autònoma de Barcelona, UAB
Specialist in psychometric evaluation.
eduardo.doval@uab.cat

M.A. Pedro Silvio Conrado González
UNAN-Managua, FAREM-Carazo
Master's Degree in Specific Didactics with Curricular Management.
pedrosilvioc081180@hotmail.com

PhD. Marta Fuentes Agustí
Universidad Autònoma de Barcelona, UAB
Specialist in teaching and learning strategies.
marta.fuentes@uab.cat

PhD. M. Dolors Riba
Universidad Autònoma de Barcelona, UAB
Specialist in statistical analysis.
dolors.riba@uab.cat

PhD. Jordi Renom
Universitat de Barcelona, UB
Specialist in psychometric evaluation.
jrenompinsach@ub.edu

DOI: https://doi.org/10.5377/torreon.v7i18.7715

Keywords: *Atypical response patterns, modified precautionary index, knowledge evaluation test, validity.*

SUMMARY

Test or test type tests provide good academic knowledge acquired by students. Although the tests are designed correctly, with contents representative of the knowledge that is to be evaluated, the results obtained by students who answer the questions atypically can be biased indicators of their levels of knowledge. This possible invalidity of some individual scores can be studied by identifying Atypical Response Patterns (ARP). However, the identification of ARP does not provide information about the causes of it.

The objective of this work is to identify some of these possible causes. For this, the answers of a same test of 136 students of three careers taught in the Regional Multidisciplinary Faculty of Carazo, of UNAN-Managua (FAREM-Carazo) have been analyzed. Twenty-six of the students answered atypically, thanks to voluntary interviews with 16 of them; it could be identified as possible explanations for the presence of ARP in the absence of study and the consequent random answers to questions considered difficult or even, in the presence of a response copy. All these reasons justify the doubt about the validity of the scores obtained by those students who gave answers to the test based on aspects different from those of their knowledge in the subject evaluated.

INTRODUCTION

In the academic context, it is frequent to perform the evaluation of knowledge through evaluation tests with a certain number of questions, the same for all the students evaluated. The so-called test-type tests, with specific and limited response options, constitute one of the most popular test formats due to their relative ease in processing and their objectivity in the correction. Teachers who develop this type of evidence are concerned, and rightly so, that the content of the same, both in terms of the statements of the questions and the response options,it is relevant to the content evaluated and that these contents are represented to the maximum by the questions and answers raised. On these aspects, the teacher has helped (Haladyna and Rodriguez, 2013, Lane, Haladyna and Raymond, 2016, Moreno, Martinez and Muñiz, 2015), and having their opinion, you can validate the content of the test to make a good evaluation. However, this does not guarantee the validity of the test, since other aspects may threaten it. One of these aspects concerns the way in which the evaluated student issues his answers.

It is assumed that a student has to answer an exam solely and exclusively based on the level of knowledge that he / she has on the subject evaluated. If so, the resulting scores are usually the sum of correct answers and can be interpreted in the expected terms: a student with a high level of knowledge will answer many questions correctly and therefore will get a high score, while otherwise, a student with low level of knowledge will not be able to answer many questions well and that will be reflected in a low score. With the same logic, it would be expected that a student who answers some questions well and badly others, does not follow any pattern of answers, but what would be expected is that he answered correctly easier questions and failed more difficult questions. This poses a dilemma, when a student answers in an illogical way from this point of view, for example guessing the most difficult questions and failing the easiest ones. In this way, two students could have the same score (for example, 5 points obtained in a test of 10 questions), but one answering well the five easiest questions and the other, the five most difficult. Differences such as the one raised, generate a series of doubts about the validity of the test scores: given that the two students have obtained the same score, can it be deduced that

they have acquired the same level of knowledge? What explains that a student who is capable of answering high level difficulty questions does not show competence when answering very easy questions?

The first question can be answered by analyzing what is known as Atypical Response Patterns, or unexpected ways of responding to the test, such as the one based on the level of difficulty of your questions. To identify the presence of PAR, numerous indexes have been developed (Karabatsos, 2003, Meijer and Sitjsma, 2001). The mere identification of ARP however, it does not allow to justify its presence, and therefore, it is necessary to inquire about the reasons with other complementary procedures, such as, for example, directly asking the students about the way they have answered the exam (Petridou and Williams, 2010), something necessary to answer the second of the questions posed.

This study aims to find an explanation for the presence of atypical patterns of response in a test of academic knowledge evaluation.

METHOD

Subjects

Out of 136 FAREM-Carazo students evaluated, 46 are first-year careers in Tourism and Hotel Management, 45 in Education Sciences with a mention in Mathematical Physics, 45 in Education Sciences with a mention in Language and Literature. All of them studied the subject Geography and History of Nicaragua that is taught in the first year of these careers. The administered test evaluated knowledge of the subjects taught in the first two units of the subject: Unit I. Introduction to the study of Geography and History of Nicaragua for citizenship and professional training and Unit II. Territorial identities and cultural identities of Nicaragua.

Instruments

The evaluation test consisted of 30 questions; 15 multiple choice, with four response alternatives, and 15 true/false response. The maximum score in the test was 25 points, which corresponded to 25% of the corresponding grade to the accumulated previous to the final exam. The other 75 % was achieved with another short test and 2 written papers.

Question 21 presented a problem and was withdrawn from the examination, so 29 questions were examined for the analysis.

Subsequent to the evaluation test, some students were interviewed. The script of the interview included the following questions.

1. What was the didactic strategy to solve the exam?
2. Do you consider that the pre-test preparation was sufficient to resolve it?

3. At the time of answering the exam, what were the main difficulties encountered in it?
4. After reading the test, did you answer any questions at random when answering?
5. After reading the exam, at the time of answer, did you copy any response from your classmates?
6. Do you consider that the extension of the exam was adequate to evaluate the contents?
7. Do you think that the time allotted to answer the test was enough to answer all the questions?

Process

The evaluation test was carried out at the time and time foreseen by the Faculty. The students' answers to the test questions were coded as hits (1) and errors (0). The individual responses to the test were analyzed in order to identify atypical patterns of response. This identification was made in two ways: calculating the Modified Caution Index (IPM) (Harnish and Linn, 1981) and comparing the profile of observed hits (O) with what would be expected according to the deterministic model (M) of Guttman (Doval and Riba, 2016, Doval, Riba, García-Rueda and Renom, 2016, Riba, Doval, Renom and Fuentes, 2017).

In both indices, the pattern reference of correct answers is Guttman's model (1950). This model determines that in a test of K questions, a person who obtains an X score (being X<K), should have correctly answered the X easiest items and answered incorrectly the K-X most difficult items.

The IPM compares the patterns of observed responses with the perfect Guttman pattern (correctly answering the K easiest questions) and the inverse Guttman pattern (correctly answering the K most difficult questions), all of which are weighted by the difficulties of the questions. It provides values between 0 (expected response pattern) and 1 (response pattern completely opposite to expected). In this study, it was calculated with the *Perfit* package from R (Tendeiro, 2015) and possible indicators of RAP and IPM values equal to or higher than 0.30 were considered (Karabatsos, 2003).

The success profile (Doval and Riba, 2016, Doval, Riba, García-Rueda and Renom, 2016) is obtained as follows. The questions are divided, according to the centile assigned to their difficulty index, into three groups: low difficulty (centile equal to or less than 33), medium difficulty (centile between 33 and 66) and high difficulty (centile higher than 66). Then, the percentage of correctly answered questions is calculated within each block. The graphic representation of these percentages is the profile of observed hits (O: see figure 1).

On the other hand, the percentage of correct answers in each block is calculated taking into account the model of Guttman (1950). Specifically, in a test of 30 questions (10 of low difficulty, 10 of medium difficulty and 10 of high difficulty), a person who has correctly answered 15 questions, according to the Guttman model should have answered the 10 easiest items (100% of the low difficulty block) and also the next 5 items in difficulty (50% of the medium difficulty block) and incorrectly answer the 10 most difficult items (0% of the high difficulty block).

The graphic representation of these percentages forms the success profile according to the model (M: see figure 1). The Euclidean distance between profiles O and M was used as an indicator of the presence of PAR (Riba, Doval, Renom and Fuentes, 2017). A Euclidean distance equal to or greater than 0.50 was considered an indicator of possible presence of PAR.

The guidelines that fulfilled the two previous criteria (IPM>0.30 and Euclidean distance>0.5) were considered RAP.

In order to deepen the reasons for the presence of RAP, students were asked to attend an individual interview with the subject's teacher voluntarily and without consequences in the result of the evaluation.

The type of RAP was identified by comparing, by difference, the observed profile (O) and the profile of the model (M), which provides a new profile (O-M) that illustrates the deviation of the observed responses with respect to those modeled. To the right of Figure 1 is shown the O-M profile resulting from comparing the O and M profiles shown to the left. The case represented as shows a relevant deviation profile in the low and medium difficulty blocks (fewer correct answers than the modeled ones) and in the high difficulty one (more correct answers than the modeled ones).

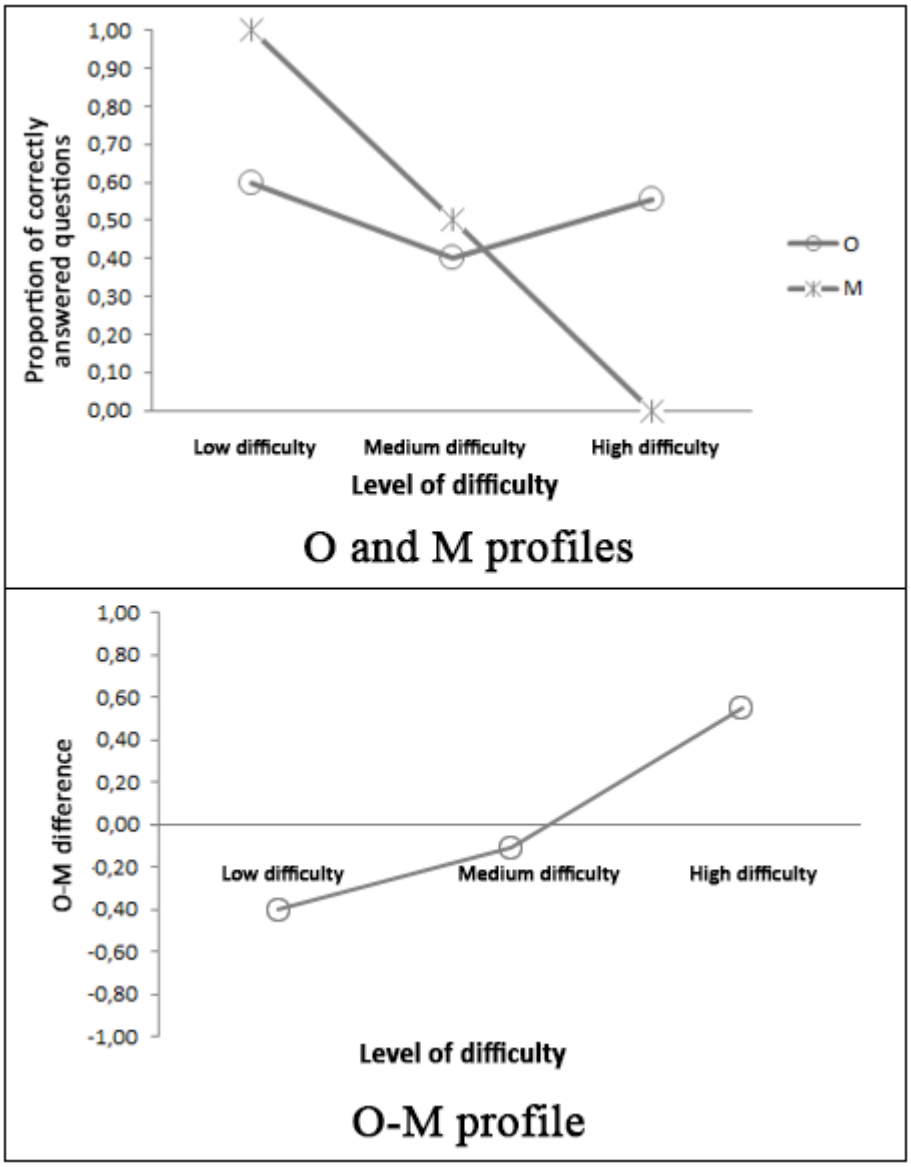


Figure 1. Profiles of observed hits (O) and modeled (M) and difference profile (O-M)

RESULTS

The difficulty profile of the exam can be seen in figure 2. The 10 easiest questions were the block of low difficulty questions, the next 10 the block of questions of medium difficulty and the 9 most difficult, the block of questions of high difficulty.

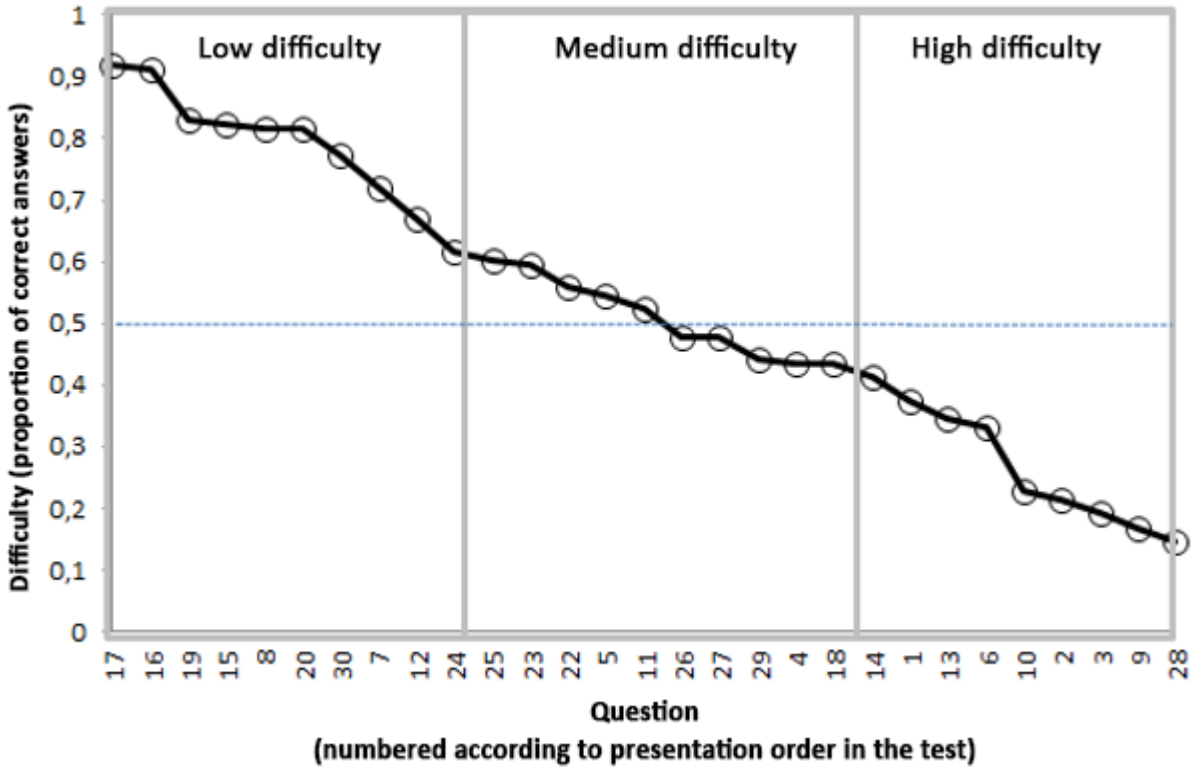


Figure 2. Difficulty profile of the test

The distribution of scores (min = 5, max = 21, M = 15.4, DE = 2.92, Asymmetry = -. 50) is shown in figure 3. The distribution of the IPM indices (M = .48, DE = .16, Asymmetry = .29) and Euclidean distance between profiles (M = .22, SD = .08, Asymmetry = .35) shows that most of the response patterns were not atypical. Twenty-eight PARs were identified, 20.6% of the total response patterns. The distribution of scores of these students is similar to that of the set (min=8, max=21, M=16, SD=3.22, Asymmetry= -.56).

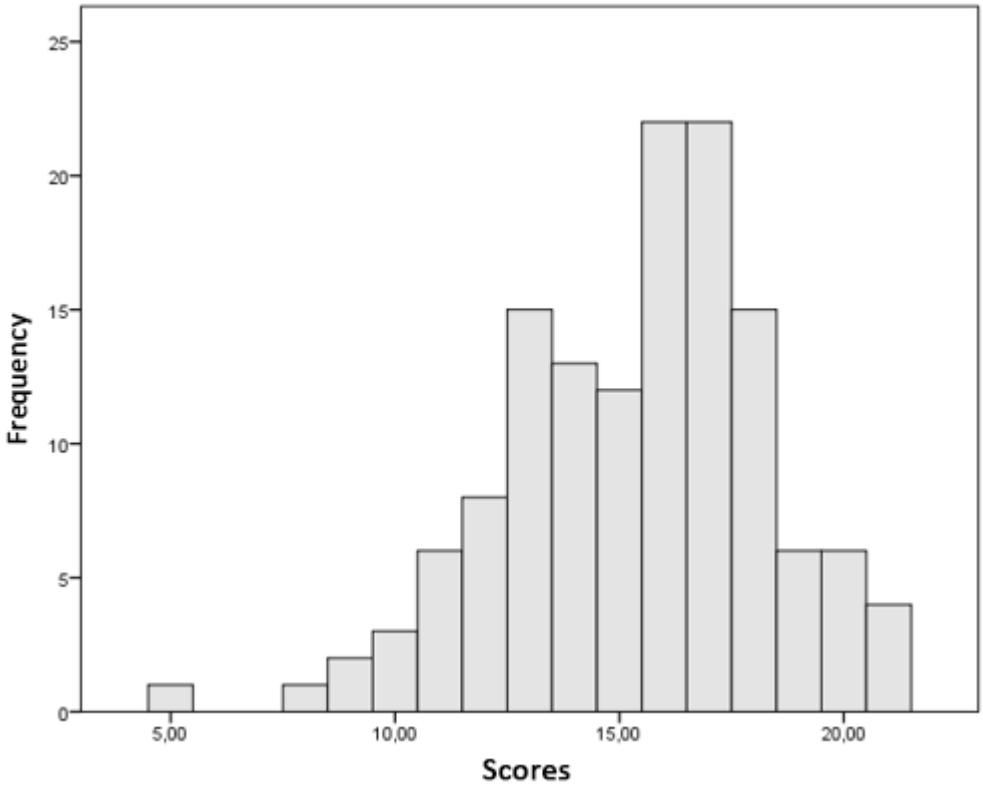


Figure 3. Distribution of test scores

Of the 28 students who answered atypically, 16 agreed to be interviewed by the teacher. Ten of these students (62.5 %) indicated that the size of the test was adequate, and the rest (37.5 %) felt that it was very long, but almost all of them (87.5 %) considered that the time had to perform the test was sufficient and only two (12.5 %) said they had to answer quickly. These two students also felt that the test was long.

Of the total students, 9 (56.25 %) indicated that they had studied little, four students (25 %) said they had studied partially, another two (12.5 %) said they had studied the necessary and only one (6.25 %) claimed to have studied a lot.

Regarding the test, seven students (43.75 %) indicated that the questions were confusing, three (18.75 %) said that the questions were difficult and another three (18.75 %), that the problem was that they had not studied enough. Two students (12.5 %) felt that the test had too many questions and only one (6.25 %) said that they considered the test to be normal in difficulty and length.

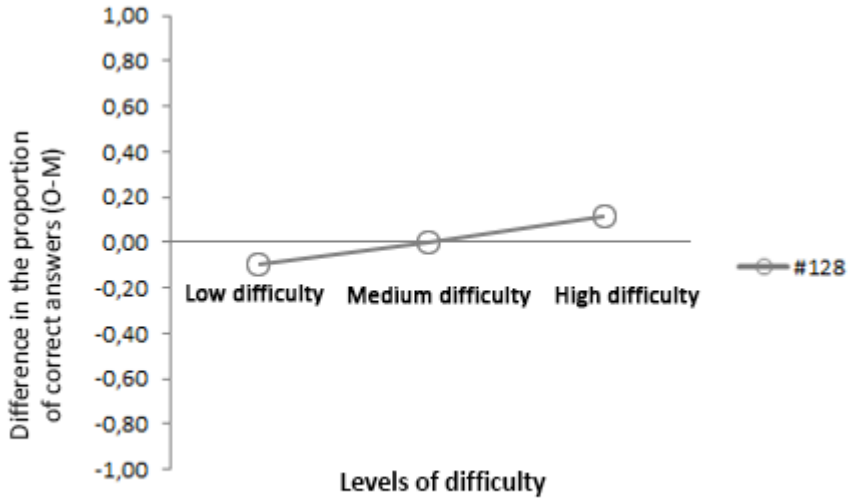
Respecting to the strategies followed to answer the questions, 9 (56.25 %) of the students answered first the questions they considered easier and left the most difficult for the end; 4 said to answer the questions running, respecting their presentation order (25 %). One student (6.25 %) indicated that their answers were based on reading and detailed analysis of the questions. Two students (12.5 %) stated that in most of the questions they looked for clues or

clues to choose the answer. More than half of the students evaluated said they answered some random questions (11: 68.75 %). Two students (12.5 %) claimed to have copied some questions.

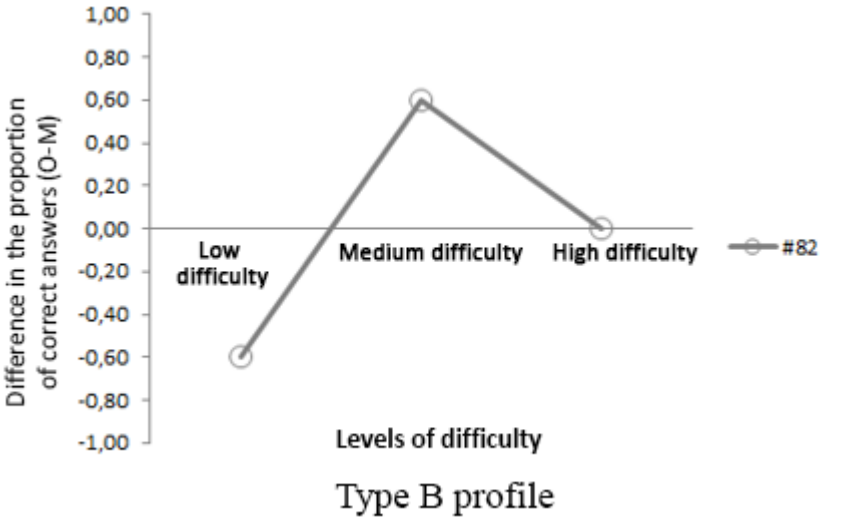
Considering all three aspects together (domain / non-domain, random / non-random answers and copy / non-copy), the most frequent profile has been that of students who claimed to have no domain and answer randomly some questions (6: 37.5 %), followed by those who said they had mastered the subject, but, even so, they had answered some questions at random (4: 25 %) and those who said they did not master the subject, although they indicated that they did not answer the questions of invalid way (3: 18.75 %). More residual, two students (12.5 %) claimed not to master the subject, and one of them said they copied some answers and another, in addition, they answered some random questions.

Figure 4 contains different examples of O-M profiles observed in the responses of the students evaluated. Profile A corresponds to a student without RAP. As can be seen, the deviation between the percentage of correct responses observed and modeled is small or zero at each of the three levels of difficulty. The rest of the profiles can be RAP identifiers, depending on the relevance of the deviations. In the type B profile, the correct answers in the block of easy questions are much less than what might be expected, while in the block of questions of medium difficulty the opposite occurs. In the C type profile, it is in the difficult questions in which there are more correct answers than what would have to be expected, while in the type D profile these unexpected correct answers are observed in the block of questions of medium and high difficulty. Finally, in the E-type profile, in the blocks of questions of low and medium difficulty, fewer correct answers than expected are observed and, nevertheless, it is in the block of more difficult answers in which they are observed, unexpectedly, more correct answers.

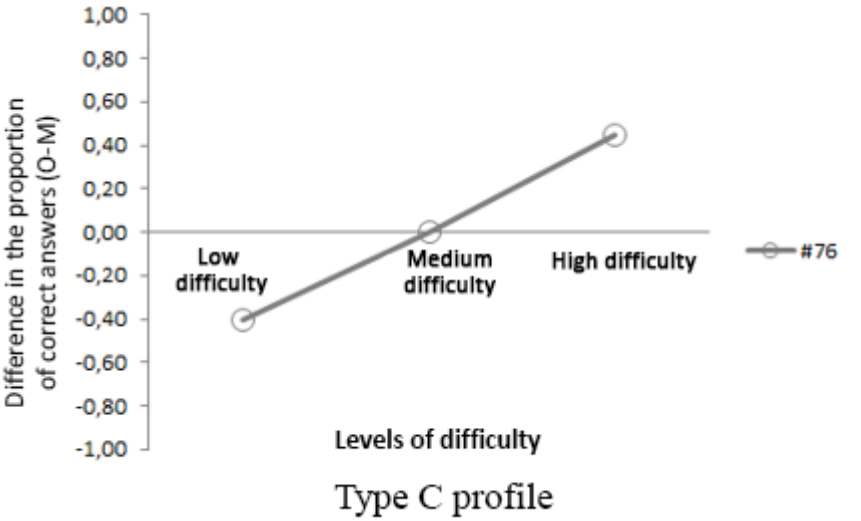
Eigthy–one. Twenty five percent (13) out of 16 students interviewed responded according to a pattern of type E responses and the rest (3: 18.75 %) with a pattern of type D responses.



Type A profile



Type B profile



Type C profile

Table continues on next page

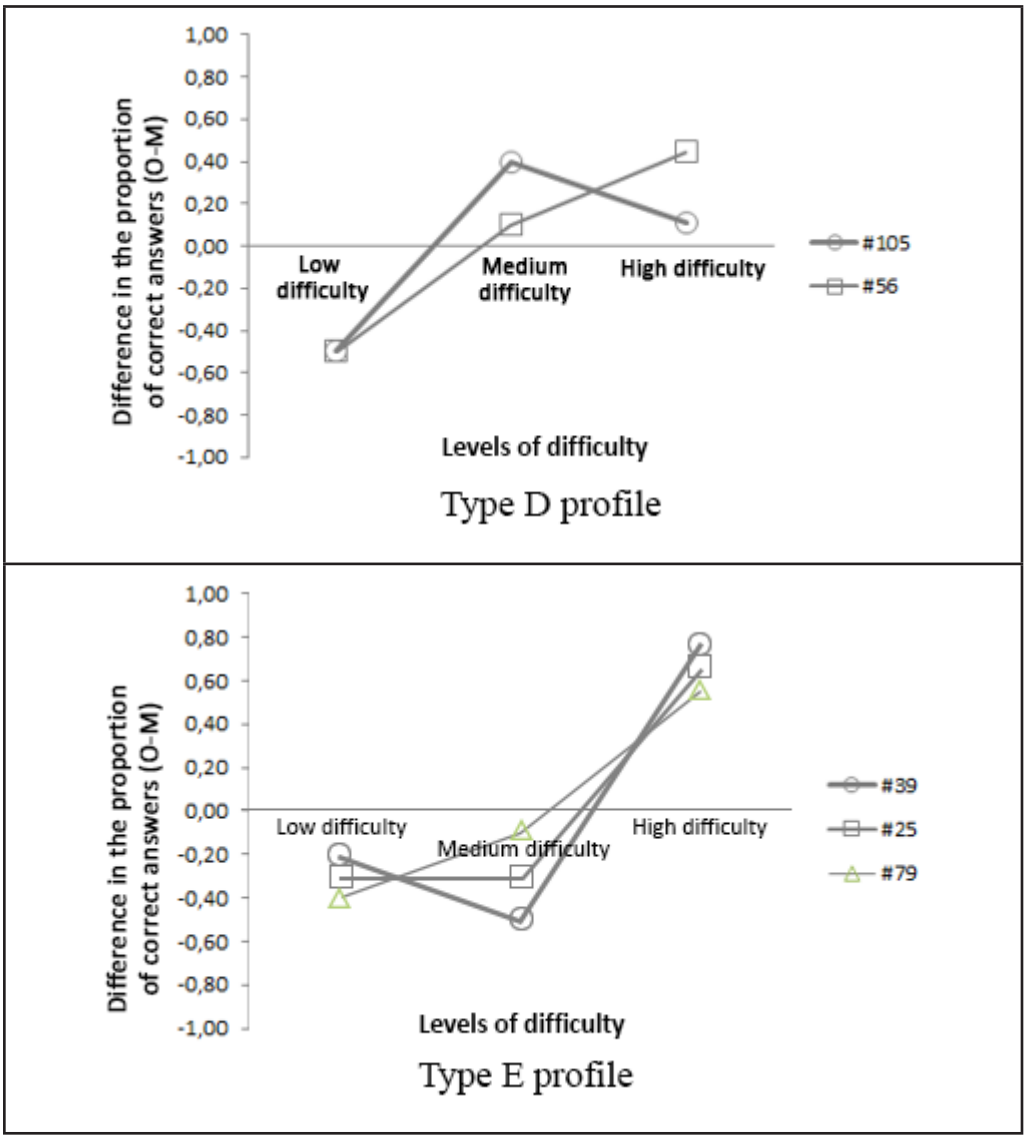


Figure 4. Different types of O-M differences. Types B, C, D and E can represent types of RAP

DISCUSSION AND CONCLUSIONS

It is significant that the majority of the students interviewed, whose response patterns had previously been identified as atypical patterns, it was stated that they had answered that they had not answered all the questions based solely on their knowledge of the subject being evaluated. In fact, most of these students also confessed that they had not prepared enough in the subject. In this context, it is not surprising that the interviewed students answered better to the more difficult questions than to the easier ones, especially when they themselves affirm that, it was given the ignorance of the correct answers, they chose mainly to answer at random and in a lesser case, for copying the answers of a partner with more knowledge.

Answering randomly is a behavior that, up to a point, it is accepted in the academic context, but copying constitutes reprehensible behavior. Perhaps, this fact justifies that the

students interviewed have confessed more answers randomly than they copied. The procedure followed, which assured the students that no reprisals would be taken in the grades, does not guarantee, however, the total confidence in the answers in the interviews. In spite of this, from the justifications gathered in the interviews, sufficient evidence has been obtained that the scores obtained by the students identified with RAP are not valid to identify the true level of knowledge acquired in the subject. And, although the presence of RAP may indicates an undervaluation or overvaluation of the level of knowledge, the reasons given by the students interviewed indicate that, in this case, the scores obtained by the vast majority of them overvalued their knowledge. Only in one case, the student said he had prepared well in the matter and considered that the test was not difficult and that the time to answer it was adequate. His score was high (19 points) and despite this, he answered in an atypical way the easy questions as he failed to answer them and instead, answered more correct questions in the block of difficult questions than of medium difficulty. If he had really studied the subject, it is possible that the problem in this case was that he was able to answer difficult questions and, therefore, the problem would be to find out why he did not correctly answer more questions of medium difficulty. This may be a case of a score that underestimates the student's true ability.

Be that as it may, the analysis carried out has made it possible to detect a group of students who have been incorrectly evaluated since the inferences, which can be made based on the scores obtained, based on the way of answering the test. They have a dubious validity.

It is considered that the validity of the scores must be guaranteed in all educational evaluations (AERA, APA, NCME, 2014) and that the described method can be very useful to identify possible sources of disability. In addition, this method is also applicable to exams with open-ended questions, since its correction in terms of correct answer or incorrect answer also defines a response pattern that can be analyzed. However, the RAP analysis is not an infallible method to identify patterns that invalidate the scores obtained. For the conclusions that can be drawn from it, it should be taken with caution, and if possible, before the teacher takes action in this regard, be sure to expand the evidence pointing to an undervaluation or overvaluation of the scores obtained in the test.

THANKS

This research has benefited from the studies carried out by financing the General Directorate of Research and Management of the National R + D + i Plan of the Ministry of Economy and Competitiveness of Spain (Project EDU2013-41399-P) and the projects of cooperation FSXXVIII, AECID Ref. 1/041892/11 and FSXXXIII.

REFERENCES

American Educational Research Association, American Psychological Association, National Council on Measurement in Education. (2014). *Standards for educational and psychologicaltesting.* Washington, DC: American Educational Research Association

Doval, E. y Riba, M.D. (2016). Identificación de tipologías de Patrones atípicos de respuesta en pruebas tipo test. *V Congreso internacional multidisciplinar de investigación educativa, CIMIE16*. Sevilla (España).

Doval, E., Riba, M.D., García-Rueda, R. y Renom, J. (2016). Comparison of the capacity of three nonparametric person -fit indexes to detect different aberrant response patterns on real data.*VII European Congress of Methodology*. Palma de Mallorca (España).

Guttman, L.A. (1950). The basis for scalogram analysis. In Stouffer, S.A., Guttman, L.A., y Schuman, E.A., *Measurement and prediction*. Volume 4 of Studies in social psychology in World War II. Princeton: Princeton University Press.

Haladyna, T.M., y Rodríguez, M.C. (2013). *Developing and validating test items*. New York, NY: Routledge

Harnisch, D. L., y Linn, R. L. (1981). Analysis of item response patterns: Questionable test data and dissimilar curriculum practices. *Journal of Educational Measurement*, 18, 133–46.

Karabatsos, G. (2003). Comparing the aberrant response detection performance of thirty-six person-fit statistics. *Applied Measurement in Education 16,* 277–298.

Lane, S., Haladyna, T.M. y Raymond, M. (2016). *Handbook of test development* (2nd Ed.). New York, NY: Routledge.

Meijer, R.R. y Sitjsma, K. (2001). Methodology Review: Evaluating Person Fit.Applied Psychological Measurement, 25 (2), pp. 107–135.

Moreno, R., Martínez, R.J. y Muñiz, J. (2015). Guidelines based on validity criteria for the development of multiple-choice items. *Psicothema*, 27(4), 388-394.

Petridou, A. y Williams, J. (2010). Accounting for unexpected test responses through examinees' and their teachers' explanations. *Assessment in Education: Principles, Policy & Practice, 17*:4, 357-382.

Riba, M.D., Doval, E., Renom, J. y Fuentes, M. (2017). Propuesta para detectar patrones atípicos de respuestas en contextos reales de evaluación. *XV Congreso de metodología de las ciencias sociales y de la salud*. Barcelona (España).

Tendeiro, J. N. (2015). Package 'Per Fit' [Software]. University of Groningen. Available at http: //cran.r-project.org/ web/packages/PerFit