Characterization of behavior of correctors when grading mathematics tests[1]

Caracterización de las actuaciones de correctores al calificar pruebas escritas de matemáticas

DOI: 10.4438/1988-592X-RE-2015-371-307

Alberto Arnal-Bailera
José María Muñoz-Escolano
Universidad de Zaragoza
Antonio M. Oller-Marcén
Centro Universitario de la Defensa de Zaragoza

Abstract

In this work, we present some results obtained from the analysis of the behavior of 91 mathematics teachers (prospective, secondary education and university) when they grade three different types of correct answers to a classical high school problem through a questionnaire. In addition to a descriptive analysis that studies the variability and the interrater reliability, we analyze the role of experience and training as well as the influence of the different solving methods. Furthermore, we try to identify profiles of correctors among secondary education teachers using both quantitative (cluster analysis) and qualitative (content analysis) methods. In particular, we observe a great variability on the assigned grades as well as a low interrater reliability. The belonging to a particular group has impact over the assigned rates while experience has no significant influence. The grades are higher when methods closer to the corrector are used. Finally, we have been able to identify three different clusters, which are determined by

[1] This work is carried out by the «S119-Investigación en Educación Matemática» research group funded by the Aragon Government and de European Social Fund. It has been partially funded thanks to the Project EDU2012-31464 (Spanish Economy and Competitiveness Ministry).

Received: 03-06-2015 Accepted: 09-10-2015

the comments and actions regarding three aspects of the students' answers: argumentation, correctness and method.

Keywords: scoring, mathematical tests, evaluators, profiles, interrater reliability.

Resumen

En este trabajo presentamos algunos resultados obtenidos al analizar el modo en que 91 profesores de matemáticas (en formación, de Secundaria y de Universidad) califican 3 tipos de respuestas correctas de un problema típico de Bachillerato a través de un cuestionario. Además de un análisis descriptivo con el que se estudia la variabilidad en las calificaciones y la fiabilidad interjueces, analizamos el papel de la experiencia docente y la formación de los correctores así como la influencia de los distintos métodos de resolución. Por otro lado, abordamos la identificación de perfiles de correctores entre los profesores de Educación Secundaria utilizando métodos cuantitativos (análisis de conglomerados) y cualitativos (análisis de contenido). En particular se observa una gran variabilidad en las puntuaciones otorgadas y una baja fiabilidad interjueces. El colectivo de pertenencia tiene impacto sobre las calificaciones de los correctores mientras que la experiencia docente no influye significativamente. La calificación otorgada por parte de los correctores es mayor cuando se utilizan métodos más cercanos a su práctica docente. Finalmente, se han identificado tres conglomerados de correctores caracterizados por sus comentarios y actuaciones relativos a tres aspectos de las respuestas de los alumnos: la argumentación, la corrección matemática y el método de resolución utilizado.

Palabras clave: calificación, exámenes de matemáticas, evaluadores, perfiles, fiabilidad entre correctores.

Introduction and background

In the Organic Law 8/2013, for the improvement of the quality of education (LOMCE) it is stablished as "one of the main novelties" to perform external assessments at the end of each educative stage. It also stablishes that these tests will have "formative and diagnostic character".

Thus, it is obvious that the result of these external assessments will be of great administrative, institutional and social importance. The legislator himself recognizes this by pointing out the necessity of "transparency" as well as that the tests must "be careful [...] in order to measure the results of the learning process".

Received: 03-06-2015 Accepted: 09-10-2015

The closest antecedent of this kind of external assessments is given by the University Entrance Examinations (Spanish acronym, P.A.U.). Even if the objectives of these new external assessments are not the same as those of the ancient P.A.U., the truth is that the results obtained in the latter also had great social importance. In fact, several works illustrate their impact on the teaching and learning of Mathematics at the last year of High School (Contreras, Ordóñez & Wilhelmi, 2010; Ruíz de Gauna, Dávila, Etxeberría & Sarasua, 2013).

In addition to their importance, the new assessments share with the P.A.U. their external and anonymous character. Hence, the P.A.U. provide an interesting framework to analyze the behavior of different correctors when they grade the answers of students in order to improve interrater reliability and, consequently, the reliability of those assessments. In this sense, works like those by Cuxart, Martí and Ferrer (1997) or by Grau, Cuxartand Martí-Recober (2002) point out the variability arising when several correctors act upon the same exam. Gairín, Muñoz and Oller (2012b; 2013) identify eight undesired phenomena detected on the actions of the correctors and suggest measures aimed at getting over these anomalies in the correction.

Once we have notes the possible variations that may produce according to which corrector acts upon a particular exam, we aim to address the following objectives:

- To study the variability among correctors and the global reliability of their marks when they grade exactly the same answer.
- To analyze the role of teaching experience and training of the corrector regarding the grading of the answers.
- To discuss the influence of different procedures or solving method of a task on its final score.
- To outline different profiles of correctors and identify their main characteristics.
- We think that a work of this kind can contribute to the development of adequate instruments of external assessment with a view to the important role that they must play.

Received: 03-06-2015 Accepted: 09-10-2015

Theoretical frawework

In the learning and teaching processes, assessment plays a fundamental role because it is the only way to know if the student has learned what has been taught and if he is prepared for the society requirements (Rico, 2006). Apart from knowing the grade of mastery achieved by the student with reference to the proposed goals, assessment serves to determine if the teaching process has been adequate for reaching these goals (Cantón & Pino-Juste, 2011).

At an overall level, there are many essays and research about educational assessment that explore the multiple perspectives associated to this concept. We have already pointed out that educational assessment can be studied depending on the internal or external nature of the evaluator. On the other hand, it is possible to study the educational assessment on the basis of its functions in the learning and teaching process or with respect to its goals and moments of implementation (Castillo, 1999) or other assessment objects, different from the student learning, as curriculum, the teachers, educational institutions (Blázquez & Lucero, 2009) or textbooks (Monterrubio & Ortega, 2012). Different instruments have been studied to facilitate the use of any assessment method such as test, oral or written exams, work presentations, task solving observation, surveys, portfolio, software…(Moral, Caballero, Rodríguez & Romero, 2009).

In the field of teaching and learning mathematics, there are some specific studies and monographs seeking to adapt both methods and instruments to the particular needs of the subject area (Giménez, 1997, Kaur & Wong, 2011). There is a clear influence between the assessment process carried out by the teacher and the way students work in the classroom (Boesen, Lithner & Palm, 2010).

Even if there are some other instruments to assess the students' learning, written tests or exams are still widely used by Secondary school and University mathematics teachers in Spain (Álvarez & Blanco, 2014; Gil, Rico & Fernández-Cano, 2002; Palacios & López-Pastor, 2013; Rochera, Remensal & Barberá, 2002) and abroad (Cárdenas, Blanco & Caballero, 2015; McMillan, Myran & Workman, 2002).

It is not usual to include the correction of mathematics exams in the teacher training process (Mollà, 1997), although this task is carried out by nearly all of the mathematics teachers. Thus, future teachers are trained

Received: 03-06-2015 Accepted: 09-10-2015

through informal conversations, debates with other students or in-service teachers or reading about other teachers' practices.

There are not many research works that study how teachers carry out the marking task in mathematics exams. There are some studies (González, Martín-Yágüez & Ortega, 1997, Mollà, 1997) that highlight the lack of objectivity in the process of correcting written mathematics tasks. Recently, Cárdenas, Gómez and Caballero (2011) point out the subjectivity of the qualification criteria when assessing problem solving task as one of the aspects perceived by prospective teachers when they reflect about their own experience as students.

There are many factors that can cause the disparity of qualifications among different competent correctors when marking exams. Watts and García (1999) note some of them in their works about the English language exam in the P.A.U. These factors are classified in three categories. Firstly, corrector errors, such us the tendency to the midpoint of the scale, the 'halo' effect, tiredness, rush, the emotional state or the number of times the corrector has found the same mistake previously. They also note environmental caused errors and task caused errors.

While recognizing the importance of these generic factors, there are others specific to the field of mathematics that have an influence in the correcting process. These factors are related to the knowledge, conceptions and beliefs about mathematics of the correctors and to the tasks and the specific answers of the students. Hence, up to six factors that influence in the corrections are presented in the works of Wang and Cai (2006) and Meier, Rich and Cady (2006). These are the teaching experience of the corrector, the educational level where this experience has been gained, the mathematical knowledge of the corrector, his beliefs about teaching and learning mathematics, the nature of the task and the answers of the students (arising bigger differences when mathematical errors are shown)

Methodological frawework

Design of the questionnaire

In order to attain the previously stated objectives, we designed a questionnaire following the methods used by similar researches. Thus, Espinosa (2005) designed a questionnaire starting from the answers given to the same problem by four Primary Education students using different solving methods and then asked a sample of prospective teachers to grade the four solutions from 1 to 10 providing their reasons. A similar method was also recently used by Fernández, Callejo and Márquez (2014), also working with prospective teachers, and by Jarero, Aparicio and Sosa (2013), with university teachers.

Based on these ideas, we chose a problem about the computation of relative extrema of a function of one real variable because it often appears in the P.A.U. tests (Ruiz de Gauna, Sarasua & García, 2011; Zamora-Pérez, 2014).

For the selection of the students' answers, we collected evidence of different methods, procedures and solving techniques used by the students on the September 2010 examinations of Mathematics II and Mathematics applied to Social Sciences at University of Zaragoza (Gairín, Muñoz & Oller, 2012a). We also revised several Secondary textbooks from different periods of time (González & Sierra, 2004).

As a result of this analysis, we designed a questionnaire where the grading of different answers was required. This first questionnaire was validated by two doctors of Mathematics Education, alien to this study, and was piloted at the end of 2012 with six Secondary teachers. After this pilot study, some aspects of the questionnaire were modified, and it was validated again by the same experts thus obtaining the final instrument. It consisted of three answers to the same problem involving the computation of critical points of a function analogue to those appearing in the P.A.U. tests of Mathematics II from the Science and Technology specialty:

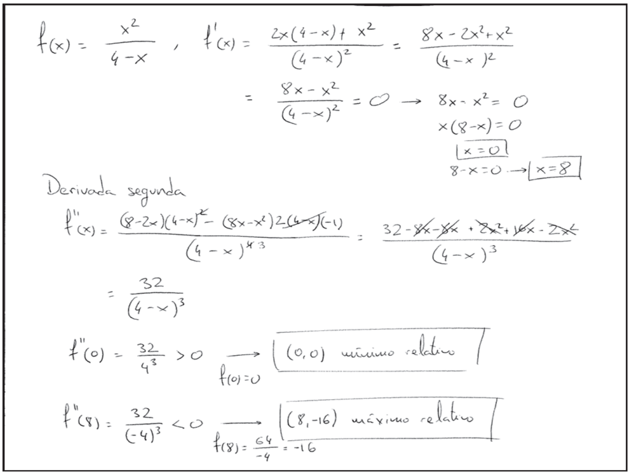
Given the function $(x) = \frac{x^2}{4 - x}$, find its relative extrema.

The three proposed answers to be graded had the following characteristics:

- The solving methods of the three answers appear in the revised textbooks and first two methods (Figures I and II) are frequently used by students in the P.A.U. tests revised.
- They do not contain any manifest Mathematical mistake.
- In the three answers, the correct solution is obtained.
- The level of argumentation (Goizueta & Planas, 2013, Yackel, 2001) of the three answers is similar. In fact, it is comparable to the mean argumentation level observed in the P.A.U. tests revised.

Below (Figures I, II and III) we show the three answers included in the final version of the questionnaire.

FIGURE I. Answer according to Method 1



$$f(x) = \frac{x^2}{4-x} \quad , \quad f'(x) = \frac{2x(4-x)+x^2}{(4-x)^2} = \frac{8x-2x^2+x^2}{(4-x)^2}$$

$$= \frac{8x-x^2}{(4-x)^2} = 0 \rightarrow 8x-x^2 = 0$$

$$x(8-x)=0$$

$$\boxed{x=0}$$

$$8-x=0 \rightarrow \boxed{x=8}$$

Derivada segunda

$$f''(x) = \frac{(8-2x)(4-x)^{\cancel{2}} - (8x-x^2)2\cancel{(4-x)}(-1)}{(4-x)^{\cancel{4}\,3}} = \frac{32-\cancel{8x}-\cancel{8x}+\cancel{2x^2}+\cancel{16x}-\cancel{2x^2}}{(4-x)^3}$$

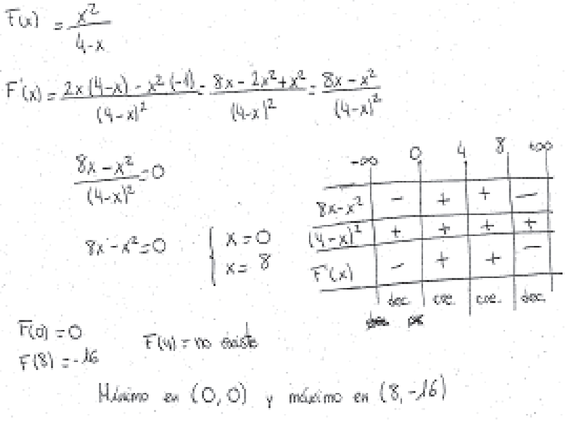
$$= \frac{32}{(4-x)^3}$$

$$f''(0) = \frac{32}{4^3} > 0 \xrightarrow[f(0)=0]{} \boxed{(0,0) \text{ mínimo relativo}}$$

$$f''(8) = \frac{32}{(-4)^3} < 0 \xrightarrow[f(8)=\frac{64}{-4}=-16]{} \boxed{(8,-16) \text{ máximo relativo}}$$

Source: Authors

FIGURE II. Answer according to Method 2.



$$F(x) = \frac{x^2}{4-x}$$

$$F'(x) = \frac{2x(4-x)-x^2(-1)}{(4-x)^2} = \frac{8x-2x^2+x^2}{(4-x)^2} = \frac{8x-x^2}{(4-x)^2}$$

$$\frac{8x-x^2}{(4-x)^2} = 0$$

$$8x-x^2=0 \quad \begin{cases} x=0 \\ x=8 \end{cases}$$

	$-\infty$ … 0	0 … 4	4 … 8	8 … $+\infty$
$8x-x^2$	−	+	+	−
$(4-x)^2$	+	+	+	+
$F'(x)$	−	+	+	−
	dec.	cre.	cre.	dec.

~~dec.~~ ~~cre.~~

$F(0) = 0$ $\quad$ $F(4)$ = no existe

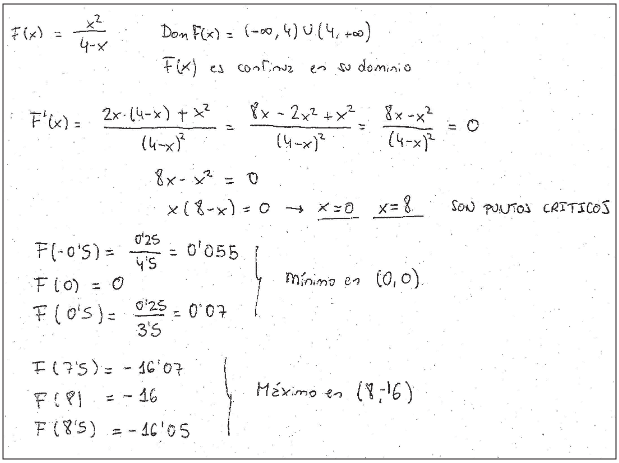
$F(8) = -16$

Mínimo en $(0,0)$ y máximo en $(8,-16)$

Source: Authors.

Received: 03-06-2015 Accepted: 09-10-2015

FIGURE III. Answer according to Method 3.



$F(x) = \frac{x^2}{4-x}$ Dom $F(x) = (-\infty, 4) \cup (4, +\infty)$

F(x) es continua en su dominio

$F'(x) = \frac{2x \cdot (4-x) + x^2}{(4-x)^2} = \frac{8x - 2x^2 + x^2}{(4-x)^2} = \frac{8x - x^2}{(4-x)^2} = 0$

$8x - x^2 = 0$

$x(8-x) = 0 \rightarrow \underline{x=0}$ $\underline{x=8}$ son puntos criticos

$F(-0'5) = \frac{0'25}{4'5} = 0'055$
$F(0) = 0$
$F(0'5) = \frac{0'25}{3'5} = 0'07$
mínimo en (0,0)

$F(7'5) = -16'07$
$F(8) = -16$
$F(8'5) = -16'05$
Máximo en (8,-16)

Source: Authors.

Together with these three answers, we included some others acting as distractors, containing some mistakes of diverse nature (Gairín et al., 2012b), with different solving methods and with an argumentation level similar to that in the three answers used for the research. In the case of Secondary teachers, in addition to the grade of each answer and the reasons for it, we asked them for their gender and the teaching experience giving lessons at the last year of High School as context variables.

Sample

91 Mathematics teachers (both in-service and prospective) have filled the questionnaire during the academic years 2012-13 and 2013-14. The sample is accidental and it is stratified according to professional categories:

- 26 prospective teachers that were enrolled in the Master's degree on the teaching of Secondary school mathematics at the University of Zaragoza (28.5% of the sample);

- 45 in-service Secondary Education teachers working on 14 high schools from Aragón and with different years of experience (49.5% of the sample);
- 20 university teachers, mathematicians that impart (or have imparted) class in Mathematics degrees (22% of the sample).

Data analysis

We perform the data analysis using a mixed research method, understood as "a set of systematic, empirical and critical processes of research involving the collection and analysis of quantitative and qualitative data as well as their integration and joint discussion" (Hernández, Fernández & Baptista, 2010, p. 546). The techniques used for the quantitative analysis are mainly statistical, while for the qualitative analysis we use mainly observational techniques (Postic & De Ketele, 1988).

Quantitative analysis

Quantitative analysis of the data focus mainly on two aspects: a descriptive statistical study, including the reliability of the whole sample and a cluster analysis (Blaikie, 2003) in the case of Secondary teachers. We restrict the cluster analysis to this stratum because the people who usually assess these contents in our educative system form this group. Cluster analysis is an easily applicable statistical technique which is little demanding regarding the characteristics of the variables. Nevertheless, it provides interesting results. For instance, some authors have used this tool to identify typologies of teachers under different criteria (Gil et al., 2002; Palacios & López-Pastor, 2013). To perform this analysis we have used the software R (version 3.0.1) and SPSS (version 15.0).

With respect to the descriptive study, we have computed the most common measures of central tendency: mean and median, as well as several indicators of the data dispersion: standard deviation, range and inter-quartile range. The comparison of means was performed by a T test.

To determine the inter-rater reliability we decided to compute the Intraclass Correlation Coefficient (ICC) of individual means, in particular, by a mixed effect model with two factors. In order to assess the degree of reliability we used the scale proposed by Fleiss (1986, p. 7).

Clusters are constructed maximizing internal cohesion and the external isolation of each group. We used k-means algorithm to construct them and Hartigan criterion (Peña, 2002) to determine the number of clusters. On the other hand, we use ANOVA techniques to observe the contribution of each variable to the existence of the clusters.

To analyze the teaching profile of each of the clusters, we study the context variables “years of experience” and “gender”: first, applying a Kolmogorov-Smirnov test to check the normality of the former and, after that, studying the differences of the means between clusters using T and U tests.

Qualitative analysis

The phase of qualitative analysis is approached by content analysis applied on the clusters identified on the previous phase. This research technique presents “many advantages and possibilities in educative and social sciences” (López, 2002, p. 177). In particular, the units of analysis are the annotations and comments written by the correctors and the different categories are constructed inductively; i.e., the categories arise from analysis itself (Berg, 2007).

Internal validity and reliability are improved with the presence of three researchers working on the same observational registers (Hernández et al., 2010, p. 476).

Results

Quantitative analysis

Descriptive analysis (global, by collectives and by questions) and reliability

Table I shows the mean, the standard deviation, the median, the inter-quartile range and the statistical range of the qualifications given by the 91 teachers that form the sample for each of the solving methods.

TABLE I. Qualifications for the three questions.

	Mean	Standard deviation	Median	$[Q_1,Q_3]$	Range
Method 1	9.53	0.84	10	[9,10]	4
Method 2	9.06	1.53	10	[9,10]	6.5
Method 3	7.87	2.14	8.5	[6,10]	8

Source: Authors.

It is interesting to note that, as the mean decreases, the standard deviation, the inter-quartile range and the statistical range increase. Moreover, this phenomenon occurs as the method becomes less «standard».

From a statistical point of view, given the results of the tests, it can be affirmed at a 99% confidence level that the mean mark given to the method 1 is higher than the given to the methods 2 and 3 and the mean mark given to the method 2 is higher than the given to the method 3.

TABLE II. Marking results of the three answers by collective.

		Mean	Standard deviation	Median	$[Q_1,Q_3]$	Range
Prospective ICC=0.178	**Method 1**	9.58	0.98	10	[9.63,10]	4
	Method 2	8.65	2.15	9.75	[9,10]	6.5
	Method 3	8.02	2.46	9.25	[6,10]	8
Secondary ICC=0.401	**Method 1**	9.51	0.86	10	[9,10]	4
	Method 2	9.03	1.33	9.5	[9,10]	5
	Method 3	7.48	1.98	8	[6,9]	6
University ICC=0.197	**Method 1**	9.51	0.59	9.88	[9,10]	2
	Method 2	9.64	0.60	10	[9.38,10]	2
	Method 3	8.55	1.93	9.5	[7,10]	6

Source: Authors.

The phenomenon pointed out in the general analysis can be observed among prospective teachers and among in-service Secondary school teachers when we take a closer look to the data by collectives (Table II). Means decrease and standard deviations increase. Differences between means are statistically significant at a 99% confidence level for in-service teachers and at a 95% for prospective teachers. Nevertheless, among University teachers methods 1 and 2 receive similar marking. In fact, there are no statistically significant differences between means and the standard deviation is almost identical. However, there were differences between the means of methods 1 and 2, at a 99% confidence level. These two were higher than the mean of the third method.

The dispersion on the samples is remarkable. Notwithstanding, this statement can be tinted because the inter-quartile ranges are lesser than one for methods 1 and 2. However, with regard to method 3, we appreciate a high dispersion since the inter-quartile ranges are bigger than 3.

In some collectives and methods (especially in training teachers for methods 2 and 3) it is observed that the data distributions are barely symmetric with a left bias. Moreover, there is a big difference between median and mean which lies out of the interval [Q1, Q3] for the second method.

No statistically significant differences can be observed between marks given by the three collectives for the first method. Regarding to method 2, we can claim (99% confidence level) that the University teachers give higher marks than prospective teachers. Even if the mean mark in the in-service Secondary school teachers is higher than the one for prospective teachers, this difference does not result statistically significant. No differences can be observed between Secondary school and University teachers. In respect of method 3, University teachers give higher marks (95% confidence level) than the other two collectives. Prospective teachers give higher marks (99% confidence level) than Secondary school teachers.

If we look at the inter-rater reliability, the value of the ICC is 0.284 for the whole sample. This indicates a poor reliability according to the Fleiss criterion (1986). Reliability at any of the strata are not acceptable either (Table II)

Establishing clusters for Secondary school teachers.

For the clusters construction we use the k-means algorithm with Euclidean distance. Regarding the number of clusters, we use Hartigan criterion (Peña, 2002), we get 65.78 and 7.52 as F-values for 2 and 3 clusters respectively. This criterion suggests the use one cluster more when the F-value is bigger than 10, so in our analysis we will use 3 clusters. A relevant number of individuals is assigned to each of the clusters: 7, 23 and 15 respectively (Table III).

TABLE III. Clusters centers and distances between them.

Clusters centers				Distances between the centers			
Cluster	Method 1	Method 2	Method 3	Cluster	1	2	3
1	8.00	6.57	6.29	1		4.298	3.846
2	9.74	9.26	9.15	2	4.298		3.732
3	9.87	9.83	5.47	3	3.846	3.732	

Source: Authors.

Applying an ANOVA (Table IV), we observe statistically significant differences in all variables at a 99% confidence level.

TABLE IV. ANOVA for the three methods.

	Cluster		Error			
	Quadratic mean	gl	Quadratic mean	Gl	F	Sig.
Method 1	9.538	2	.325	42	29.309	.000
Method 2	26.609	2	.595	42	44.734	.000
Method 3	67.549	2	.902	42	74.898	.000

Source: Authors.

Once we have justified the number of clusters and the relations among them, we show and analyze the descriptive statistics within each cluster (Table V).

TABLE V. Descriptive statistics for each cluster.

Cluster	Method	Mean	Standard deviation	Min	Max	Teaching experience mean	Teaching experience Std. deviation	% women
C1 (N=7)	M1	8.0000	1.15470	6.00	9.00	17.00	7.38	43.86%
	M2	6.5714	1.13389	5.00	8.00			
	M3	6.2857	1.11270	5.00	8.00			
C2 (N=23)	M1	9.7391	.42291	8.50	10.00	14.41	10.40	52.17%
	M2	9.2609	.78146	7.00	10.00			
	M3	9.1522	.76030	8.00	10.00			
C3 (N=15)	M1	9.8667	.35187	9.00	10.00	10.60	8.89	60.00%
	M2	9.8333	.52327	8.00	10.00			
	M3	5.4667	1.12546	4.00	7.00			

Source: Authors.

Method 1 receives, on average, marks equal or higher than 8 in the three clusters. Even if marking is very high in all the cases, we observe that teachers in clusters two and three consider this method virtually perfect whereas the teachers in cluster one give two points less on average.

Method 2 receives, on average, marks lower than 6.6 points by the teachers in the first cluster, whereas this exercise has been marked with more than 9.2 points in the other two clusters. We can claim that method 2 is not totally accepted by teachers in the first cluster.

Method 3 is marked, on average, in a different way in each of the three clusters, getting a mean over a 9 in the second and 6.29 and 5.47 in the other two. We can affirm that method 3 is totally accepted only by teachers in the second cluster.

Received: 03-06-2015 Accepted: 09-10-2015

Considering the variability in the answers, we observed very high standard deviations in the whole set of teachers. Now, we see how these standard deviations keep high –especially in the first cluster– being even higher when marks are lower. This points to a different marking of the errors by each teacher.

We can partially characterize a certain type of teacher when facing the marking of mathematically correct exercises:

- Cluster 1 is comprised by teachers that grant low marks on methods 2 and 3. They only consider totally correct the first method.
- Cluster 2 is comprised by teachers that grant high marks on the three methods. More than half of the teachers on the sample have been assigned to this cluster
- Cluster 3 is comprised by teachers that grant low marks on method 3 and high marks on methods 1 and 2. It includes one third of the teachers on the sample.

In order to analyze if there is a relation between gender, teaching experience in high school (Table V) and the cluster assigned, we study the statistical significance of the mean differences of these variables for each cluster. In the case of «gender», the differences are not statistically significant. «Teaching experience in high school» can be considered normal (p=0.370) but none of the differences are statistically significant using the T test. However, the Mann-Whitney test gives a statistically significant difference (90%) between clusters 1 and 3.

Qualitative analysis

As a consequence of the quantitative analysis of the data, we have identified three different groups of Secondary school teachers depending on the marks given to the three methods. We checked that these clusters cannot be totally characterized by teaching experience or gender. We now apply a qualitative analysis looking for evidences of the marking disparities among these clusters and for coincidences within them. Thus we may define the profile of the teachers and explain their reasons to mark in a particular way.

Based on successive revisions, three different topics emerge from the correctors' comments. These topics become analysis categories that we introduce hereafter with some examples for the sake of clarity (Table VI).

TABLE VI. Categories for the qualitative analysis.

Category	Description	Example
Argumentation	The corrector comments on the explanations and reasoning given by the student in his answer.	*Calcula los extremos pero no justifica ni dominio ni lo que va haciendo ni porque.* *[He calculates the extrema, but he does not justify the domain, nor his actions nor why.]* *Debería al menos justificar el hecho de que sean mínimos y máximos relativos.* *[He should, at least, justify the fact of being relative maxima and minima.]*
Mathematical correctness	The corrector comments on the mathematical correctness of the students answer.	*Un razonamiento inconsistente. Lo que hace le merece el 5. El final nada.* *[An inconsistent reasoning. Given what he does, he deserves a 5. The end, nothing.]* *Razonamiento incorrecto para comprobar si los puntos críticos son máximos o mínimos.* *[An incorrect reasoning to check if the critical points are maxima or minima.]*
Method	The corrector comments about the procedure followed by the student stating that this is unexpected for him.	*No discute el signo de la primera derivada para luego poder determinar los extremos.* *[He does not discuss the sign of the first derivative to determine the extrema.]* *Aunque no me gusta esta forma de ver los máximos o mínimos… creo que puede ser correcto.* *[Even if I don`t like this way of finding the maxima or minima... I think it can be correct.]*

Source: Authors.

Now, we are going to give consideration to what is said in each of the clusters about each of the categories of analysis in the correction protocols.

Cluster I

Comments made by these teachers in the corrections of the three methods are characterized by a constant demand of further argumentation of the students' processes. Expressions used by the teachers in methods 1 and 2 point to different requirements, some of them say that *resultados teóricos* [*theoretical results*] are needed to support the resolution and others penalize in the same manner noting that *no comenta* [*the student does not comment*]. Exigencies rise when marking method 3, where all the teachers in this cluster ask for *justification* [*justificación*] of the determination of local minima and maxima.

There are teachers that remark and penalize some partial mathematical incorrectness in the three methods. The most frequent is the lack of an explicit study of the domain in methods 1 and 2. Only two teachers consider the third method as globally incorrect: *No hace el estudio del crecimiento y decrecimiento de forma correcta* [*He does not study in a correct way the increasing and decreasing*].

Two correctors penalize the use of method 3.These teachers expressed qualification criteria by splitting down methods 1 and 2 into steps and assigning points to each one: *Obtiene los extremos relativos sin utilizar f''(x) y con un método poco fiable (1 punto sobre 4).* [*He gets the local extrema without using f''(x) and with an unreliable method (1 point over 4).*]

Teachers assigned to this cluster consider the three methods mathematically correct. Nevertheless, in the second and the third they require further argumentation, showing their preference for the first. The penalization by lack of argumentation is between one and two points in method 1 raising up to three points in the third method. Partial mathematical incorrectness, such as the lack of an explicit study of the domain in methods 1 and 2, is penalized with one point. Moreover, there are teachers who find incorrect the third method, considering it as unreliable or incomplete. This all can be seen numerically in the low qualifications given in average to methods 2 and 3.

Thus, this cluster is characterized by a high argumentation exigency - different depending on the solving method- and by the penalization of the solving methods if the function domain is not explicitly written, which is considered as a partial mathematical incorrectness.

Cluster 2

With regard to argumentation, some correctors point out a lack of *justificación o explicación [justification or explanation]*. The justification demand is more frequent on method 3. Furthermore, the penalization is higher here and some correctors do not take into account a lack of justication in method 1 because *está claro y con orden* [*it's clear and ordered*] whereas it is penalized in the third method. Another corrector asks for *razones* [*reasons*] why a point is minimum or maximum in methods 2 and 3, while he does not ask for them in the first method.

Correctors in this cluster have no doubt about the mathematical correctness of the three methods. There are hardly any objections about the study of the domain, and if there is any, this is barely penalized. A corrector says that in the first method the domain and the continuity of f (x) and f'(x) are absent, but it is very little penalized: *¿Dominio y continuidad de f(x)? ¿continuidad de f'(x)?*[*Domain and continuity of f(x)?Continuity of f'(x)?*]. There is a corrector that penalizes the lack of study of continuity in method 2 but do not ask for this study in the first one: *No justifica el uso de este método con la continuidad de la función.* [*He doesn't justify the use of this method with the function's continuity*].

Even if it is not very common in this cluster, the use of method 3 is sometimes slightly penalized with references to the higher correctness of methods 1 and 2:*El método es bueno, pero quizás hubiera sido más correcto que la comprobación la hiciera con el crecimiento o la segunda derivada*.[*The method is good, but maybe it would have been more correct if the checking would have been done using the growth or the second derivative*].Some other teachers explicitly express their personal preference for the more usual methods: *Deriva bien y obtiene los posibles extremos, pero no usa los criterios usuales para estudiarlos*. [*He derives well and obtains the possible extrema, but he does not use the usual criteria to study them*].

Teachers assigned to this cluster seem to find the level of argumentation adequate since they mark the three methods with almost 10 points. However, we observe many demands of explanation which are rarely penalized, and if so, only with half a point in methods 1 and 2 or one point in method 3.This shows a certain preference for the first two. These teachers consider the three methods mathematically correct.

This cluster is characterized by considering adequate the argumentation and the mathematical correctness of the three methods. Notwithstanding, we notice a correcting bias that points to a preference for methods 1 and 2, not concreting in a high penalization of method 3.

Received: 03-06-2015 Accepted: 09-10-2015

Cluster 3

Most of the teachers in this cluster do not require argumentation for any method. A few of them ask for it in the case of method 3, with the remarkable case of a corrector that even asks for a theorem: *No justifica con un teorema. Falta la justificación de los puntos críticos.* [*He does not justify with a theorem. The justification of the critical points is lacking*.]. This corrector do not make any request in this sense for the other two methods.

There are some teachers in this cluster that consider method 3 a mathematically incorrect procedure, finding very different penalizations because of that. Some teachers give a high mark even if they note the mathematical incorrectness: *Da un resultado correcto mediante un razonamiento erróneo.* [*He gives a correct result with an incorrect reasoning*].Some other teachers give a much lower mark motivated in a similar way. *Razonamiento incorrecto para comprobar si los puntos críticos son máximos o mínimos.* [*This is an incorrect reasoning to check if the critical points are maxima or minima*].

A significant part of the teachers in this cluster requires the explicit use of the second derivative in method 3 or a sign table to classify critical points: *O estudio del crecimiento-decrecimiento o estudio de la segunda derivada.*[*Either the study of the increase-decrease or study of the second derivative is needed*]. Some of them conclude that the student forgot or did not study the whole solving process: *Sabe que hay que derivar e igualar a cero pero no sabe y no memoriza el resto del algoritmo.*[*He knows that he has to derivate and set equal to zero but he can't and he doesn't memorize the rest of the algorithm*].

Teachers assigned to this cluster find adequate the level of argumentation and the mathematical correctness in methods 1 and 2, being very critical in both aspects with regard to method 3. Penalizations in the third method are around 5 points even if some teachers base it on the argumentation. Others do not conceive the mathematical correctness of a method different from the ones they commonly use. In some occasions, the achievement of a correct result balances the identified mathematical incorrectness.

Ultimately, the correctors in this cluster are characterized by considering completely adequate the argumentation and mathematical correctness of methods 1 and 2, with very few comments at this respect.

They show serious concerns about the appropriateness of method 3, even rejecting it explicitly or penalizing aspects not penalized in methods 1 and 2.

Discussion

The task of correcting and grading problems is far from being easy and it requires a deep thought about the contents (concepts, procedures) that are evaluated about the different strategies that can be used to solve them (Gairín et al. 2012b, 2013). In this work, we have checked that even for tasks where a broad consensus about their correction exists, it is not absolute. Regarding objective 1, it is evident the high variability among the grades assigned by the teachers, mainly for methods 2 and 3. In addition, inter-rater reliability is low both globally and within each stratum.

Regarding objective 2, our results seem to point out that the belonging to a particular group (and hence the training) has an influence on grading, being the group of prospective teachers the one with a higher dispersion on their grades. This implies that specific training about this topic might be needed, as other studies already stated (Huitrado & Climent, 2013; Mollà, 1997). Nevertheless, with respect to teaching experience, we have found some differences between the means that are not statistically significate. In any case, our findings seem to imply that with bigger samples we would have a higher amount of experienced teacher in cluster 1 versus cluster 3. These conclusions go partly in the same direction as previous studies (Meier et al., 2006; Wang & Cai, 2006) who pointed out mathematical knowledge and teaching experience as factors influencing on the variability of the grading among correctors.

Based on previous research (Gairin et al, 2012a, 2012b, 2013) and from the point of view of their presence in textbooks, we observe that method 1 and 2 are far more common than method 3. From the point of view of their use by students in the P.A.U. tests, method 3 is practically absent and method 1 is more common than method 2. Thus, our results about objective 3 indicate that the grading assigned by many of the correctors is higher when the student uses methods that are closer to the practice and teaching experience of the corrector. This finding goes on the lines of Espinosa (2005) and Fernández et al. (2014). Nevertheless, the qualitative analysis of the behavior of these correctors shows that in many cases, they do not make this preference explicit and they offer different explanations. Some correctors indicate that the chose method is not "mathematically correct", which is not the case, perhaps showing a shortage of mathematical training. Other correctors, even if they admit the correction of the method, require a higher level of argumentation in the student's answer that is not required to other students that use the method expected by the corrector.

Finally, regarding objective 4, we have identified three groups among Secondary school teachers based on their comments to three categories: the argumentation used by the student, the Mathematical correction of the answer and the concordance between the used method and the expected one. The first group is characterized by a high requirement of argumentation and the lower global grading. The second group considers argumentation and mathematical correction of the three methods to be adequate, but shows some bias in favor of method 3. The third group is characterized by a clear penalization of method 3, giving higher grades to the other two answers.

Bibliographic References

Álvarez, M.R. & Blanco, L. (2014). Sobre la evaluación en Matemáticas en Secundaria. *Revista SUMA 76*, 47-54.

Berg, B.L. (2007). *Qualitative research methods for the social sciences*. Boston: Allyn and Bacon.

Blaikie, N. (2003). *Analyzing quantitative data*.London: SAGE.

Blázquez, F. & Lucero, M. (2009). La evaluación en educación. In A. Medina & F. Salvador (Coords.) *Didáctica General* (243-291). Madrid: Pearson.

Boesen, J.; Lithner, J. & Palm, T. (2010).The relation between types of assessment tasks and the mathematical reasoning students use.*Educational Studies in Mathematics, 7*(1), 89-105.

Cantón, I. & Pino-Juste, M. (2011). *Diseño y desarrollo del currículum*. Madrid: Alianza Editorial.

Cárdenas, J.A.; Blanco, L. & Caballero, A. (2015). *Las pruebas escritas que se proponen para evaluar matemáticas en secundaria actualmente.* In XIV Conferencia Interamericana de Educación Matemática. Chiapas, México. Retrieved from: http://xiv.ciaem-iacme.org/index.php/xiv_ciaem/xiv_ciaem/paper/viewFile/231/132

Cárdenas, J.A.; Gómez, R. & Caballero, A. (2011). *Algunas diferencias entre la práctica y la teoría al evaluar la resolución de problemas en Matemáticas.* In García (Comp.), Memorias del 12º encuentro colombiano de matemática educativa (53-62). Bogotá, Colombia.

Castillo, S. (1999). Sentido educativo de la evaluación en la Educación Secundaria. *Educación XX1, 2*, 65-96.

Contreras, Á.; Ordóñez, L. & Wilhelmi, M.R. (2010). Influencia de las Pruebas de Acceso a la Universidad en la enseñanza de la integral definida en el Bachillerato. *Enseñanza de las Ciencias, 28*(3), 367-384.

Cuxart, A., Martí, M. & Ferrer, F. (1997). Algunos factores que inciden en el rendimiento y la evaluación en los alumnos de las pruebas de aptitud de acceso a la universidad (PAAU). *Revista de Educación, 314*, 63-88.

Espinosa, E. (2005). *Tipologías de Resolutores de Problemas de Álgebra Elemental y Creencias sobre la Evaluación con Profesores en Formación Inicial*. (Tesis doctoral inédita). Universidad de Granada, Granada.

Fernández, C., Callejo, M.L. & Márquez, M. (2014) Conocimiento de los estudiantes para maestro cuando interpretan respuestas de estudiantes de primaria a problemas de división-medida. *Enseñanza de las Ciencias, 32*(3), 407-424.

Gairín, J.M., Muñoz, J.M. & Oller, A.M. (2012a).*Sobre la calificación de los exámenes de las Pruebas de Acceso a la Universidad de las asignaturas Matemáticas II y Matemáticas Aplicadas a las CCSS*. Informe para la Comisión Organizadora de la Prueba de Acceso a las Enseñanzas Universitarias de Grado de la Universidad de Zaragoza.

– (2012b). Propuesta de un modelo para la calificación de exámenes de matemáticas. In A. Estepa, Á. Contreras, J. Deulofeu, M.C. Penalva, F.J. García & L. Ordóñez (Eds.) *Investigación en Educación Matemática XVI* (261-274). Jaén: SEIEM.

–(2013). Anomalías en los procesos de identificación de errores en las pruebas escritas de matemáticas de las P.A.U. *Campo abierto: Revista de Educación 32*(2), 27-50.

Gil, F., Rico, L., & Fernández-Cano, A. (2002). Concepciones y creencias del profesorado de secundaria sobre la evaluación en matemáticas. *Revista de Investigación Educativa, 20*(1), 47-75.

Giménez, J. (1997). *Evaluación en matemáticas. Una integración de perspectivas*. Madrid: Síntesis.

Goizueta, M. & Planas, N. (2013). Temas emergentes del análisis de interpretaciones del profesorado sobre la argumentación en clase de matemáticas. *Enseñanza de las Ciencias, 31*(1), 61-78

González, S., Martín-Yagüez, M.C. & Ortega, T. (1997). Propuesta y análisis de una prueba de evaluación. *Uno, 11*, 55-78.

González, M.T. & Sierra, A, M. (2004) Metodología de análisis de libros de texto de matemáticas. Los puntos críticos en la Enseñanza Secundaria en España durante el siglo XX. *Enseñanza de las Ciencias, 22*(3), 389-408.

Grau, R., Cuxart, A. & Martí-Recober, M. (2002). La calidad en el proceso de corrección de las Pruebas de Acceso a la Universidad: variabilidad y factores. *Revista de Investigación Educativa, 20*(1), 209-223.

Hernández, R., Fernández, C. & Baptista, M.P. (2010) *Metodología de la Investigación*. México: McGraw Hill Educación.

Huitrado, J.L. & Climent, N. (2013). Conocimiento del profesor en la interpretación de errores de los alumnos en álgebra. *PNA*, *8*(2), 75-86.

Jarero, M., Aparicio, E., & Sosa, L. (2013). Pruebas escritas como estrategias de evaluación de aprendizajes matemáticos: un estudio de caso a nivel superior. *RELIME, 16*(2), 213-243.

Kaur, B. & Wong, K.Y. (Eds.) (2011). *Assessment in the mathematics classroom. Yearbook 2011, Association of Mathematics Educators*. Hackensack, NJ: World Scientific.

López, F. (2002).El análisis de contenido como método de investigación. *XXI. Revista de Educación*, 4, 167-179.

McMillan, J.H., Myran, S., & Workman, D.(2002). Elementary teachers' classroom assessment and grading practices. *The Journal of Educational Research, 95*(4), 203-213.

Meier, S.L., Rich, B.S. & Cady, J. (2006) Teachers' use of rubrics to score non-traditional tasks: factors related to discrepancies in scoring. *Assessment in Education: Principles, Policy & Practice, 13*(01), 69-95.

Mollà, A. (1997). Una experiencia de formación del profesorado en evaluación en el área de matemáticas. *Uno, 11*, 79-90.

Monterrubio, M.C. & Ortega, T. (2012). Creación y aplicación de un modelo de valoración de textos escolares matemáticos en Educación Secundaria. *Revista de educación, 358,* 471-496.

Moral, C.; Caballero, K.; Rodríguez, M. J. & Romero, M. A. (2009). La evaluación en la enseñanza. In C. Moral &M. P. Pérez (Coords.) *Didáctica. Teoría y práctica de la enseñanza* (304-319). Madrid: Pirámide.

Palacios, A. & López-Pastor, V.M. (2013). Haz lo que yo digo pero no lo que yo hago: Sistemas de evaluación del alumnado en la formación inicial del profesorado. *Revista de Educación, 361,* 279-305.

Peña, D. (2002). *Análisis de datos multivariantes.* Madrid: McGraw-Hill.

Rico, L. (2006). Marco teórico de evaluación en PISA sobre matemáticas y resolución de problemas. *Revista de Educación, extraordinario 2006,* 275-294.

Rochera, M.J.; Remesal, A. & Barberá, E. (2002). El punto de vista del profesorado de educación primaria y educación secundaria obligatoria sobre las prácticas de evaluación del aprendizaje matemático: un análisis comparativo. *Revista de Educación, 327,* 249-266.

Ruíz de Gauna, J., Dávila, P., Etxeberría, J. & Sarasua, J. (2013). Pruebas de selectividad en Matemáticas en la UPV-EHU. Resultados y opiniones de los profesores. *Revista de educación, 362,* 217-246.

Ruíz de Gauna, J.; Sarasua, J. & García, J.M. (2011). Una tipología y clasificación de los ejercicios de matemáticas de selectividad. *Epsilon, 28(2), 78,* 21-38.

Postic, M. & De Ketele, J.M. (1988). *Observer les situations éducatives.* Paris: Presses Universitaires de France.

Yackel, E. (2001). Explanation, justification and argumentation in mathematics classrooms. In M. Van den Heuvel-Panhuizen (Ed.), *Proceedings of the 25th conference of the international group for the psychology of mathematics education PME-25*, vol. 1, (1–9). Utrecht (Holanda).

Wang, N. & Cai, J. (2006). An investigation of factors influencing teachers' scoring student responses to mathematics constructed-response assessment tasks. In Novotná, J., Moraová, H., Krátká, M. & Stehlíková, N. (Eds.). *Proceedings 30th Conference of the International Group for the Psychology of Mathematics Education*, Vol. 5, (369-376). Prague: PME.

Received: 03-06-2015 Accepted: 09-10-2015

Watts, F. & García, A. (1999). Control de calidad en la calificación de las pruebas de inglés de Selectividad. *Aula abierta, 73,* 173-190.

Zamora-Pérez, R.F. (2014). *Análisis de las pruebas de acceso a las universidades de Castilla y León (Matemáticas II)*. (Tesis doctoral inédita). Universidad de Valladolid, Valladolid.

Contact Address: Alberto Arnal Baileira, Universidad de Zaragoza, Facultad de Educación, Departamento de Matemáticas, Área de Didáctica de la Matemática. C/ Pedro cervuna, 12, 50009, Zaragoza, España. E-mail: albarnal@unizar.es

Received: 03-06-2015 Accepted: 09-10-2015