VARIABILIDAD EN LA CORRECCIÓN DE PRUEBAS DE MATEMÁTICAS

Elena Mengual, Núria Gorgorió, Lluís Albarracín
Universitat Autònoma de Barcelona (España)
emengualbre@gmail.com; nuria.gorgorio@uab.cat; lluis.albarracin@uab.cat

Palabras clave: Evaluación, calificación, matemáticas
Keywords: Assessment, marking, mathematics

RESUMEN

En esta comunicación presentamos un refinamiento de los criterios para la calificación de exámenes de matemáticas. Dichos criterios se redactaron a partir de la aportación teórica de Gairín, Muñoz y Oller (2012) a partir de considerar los fenómenos detectados por Mengual, Gorgorió y Albarracín (2013). El análisis de las calificaciones dadas por los correctores participantes en nuestro estudio muestra que el modelo de corrección propuesto genera una disminución de la variabilidad de las calificaciones en un porcentaje elevado de las respuestas que presentan mayores dificultades de corrección.

ABSTRACT

In this communication we present a refinement of the criteria established for the grading of mathematics exams. The latter follow the theoretical contribution of Gairín, Muñoz and Oller (2012), while observing the phenomena detected by Mengual, Gorgorió and Albarracín (2013). The analysis of the qualifications provided by graders in this study shows the proposed grading model to generate a decrease in variability of marks in a large percentage of those answers which happen to pose the most correction difficulties.

Introducción

La preocupación sobre el proceso de acceso a la universidad y las pruebas que lo conforman se hace patente desde hace años debido, en gran parte, a su influencia social. Cuxart, Martí-Recober y Ferrer (1997) señalaron la necesidad de crear una línea de investigación sobre la variabilidad de este tipo de pruebas que se muestra mayor de lo esperable.

Hasta el presente, el acceso a los estudios universitarios en España ha dependido de la superación por parte de los estudiantes de las Pruebas de Acceso a la Universidad (PAU en adelante), conocidas como selectividad, entre las que se incluye una prueba de matemáticas. El objetivo de estas pruebas es valorar la madurez académica de los estudiantes y garantizar la equidad del sistema de acceso a la universidad pública. En un futuro próximo las PAU se verán modificadas pero se mantendrá la necesidad de efectuar pruebas concretas que certifiquen los conocimientos de los alumnos en diferentes disciplinas. En concreto, la nueva reforma LOMCE establece una reválida del bachillerato y permite a los campus universitarios hacer sus propias pruebas. Aún se desconoce cómo será el nuevo acceso a la Universidad, pero lo que sí parece cierto es que seguirá habiendo un sistema que regule la entrada de los estudiantes a la universidad a partir de pruebas objetivas en España.

Las pruebas de acceso son un recurso utilizado por distintas universidades en todo el mundo. En particular, en algunos países de Latinoamérica, esta prueba es común para todos los estudiantes. Por ejemplo, en el caso de Chile algunas universidades llevan a acabo la prueba PSU (Prueba de Selección Universitaria). En Colombia todos los estudiantes de educación media deben pasar un examen obligatorio si quieren acceder a la educación superior. Mientras que en México existen dos pruebas importantes, una de ellas la realizan los estuantes de secundaria como paso previo a la entrada en la universidad. En esta dirección, en Ecuador recientemente se han creado unas pruebas de acceso a la universidad donde, entre otro temas, se evalúa la aptitud numérica y el razonamiento abstracto. En otros países no existe un examen común sino que cada universidad elabora sus propias pruebas, como el caso de Argentina, Uruguay, Perú y Venezuela. Dada la relevancia de este tipo de pruebas, consideramos que el refinamiento del modelo de calificación puede aplicarse en todos estos tipos de pruebas diversas con el fin de mejorar la variabilidad entre las correcciones otorgadas por distintos profesores a una misma prueba de matemáticas.

Nuestro estudio se centra en la prueba de matemáticas de las PAU en España. Dicha prueba constituye una evaluación sumativa y nomotética criterial, ya que la valoración del estudiante viene dada por unos criterios de corrección que elaboran los armonizadores de las PAU. Estos criterios deberían permitir valorar de forma homogénea el grado en que el alumno (anónimo para los correctores) ha alcanzado los conocimientos, destrezas y habilidades matemáticas al final de la etapa de educación secundaria. Es conocido que la corrección y calificación de las pruebas de matemáticas de las PAU no son tareas triviales, en especial en aquellos casos en los que las actividades a corregir posibilitan diferentes tipos de respuestas o los errores que pueden cometer los alumnos son difíciles de predecir. En particular, se han observado algunas dificultades en el proceso de corrección, como el hecho que los correctores no corrigen igual la primera que la última prueba (Casanova, 1998).

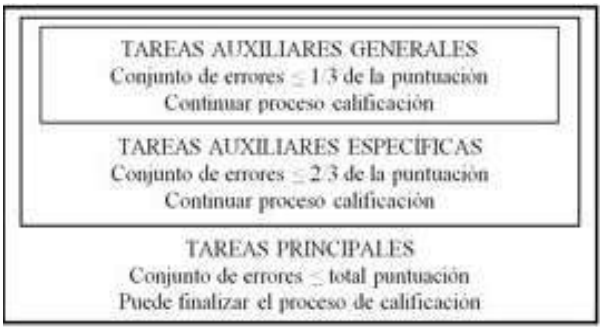
Recientemente, Gairín, Muñoz y Oller (2012) constataron la existencia de diferencias significativas entre las puntuaciones de distintos correctores cuando califican un mismo examen. Desde esta perspectiva, estos autores presentan un modelo teórico para la elaboración de unos criterios de corrección que deberían permitir una reducción en la variabilidad de las calificaciones otorgadas por los correctores. Este modelo ha sido concretado y analizado en una investigación empírica a partir de la elaboración de una guía de corrección y del escrutinio de las correcciones realizadas por un grupo de 4 profesores (Mengual, Gorgorió y Albarracín, 2013). Los resultados de dicho estudio muestran una reducción de la variabilidad entre las calificaciones otorgadas por los correctores a la misma colección de respuestas de diferentes alumnos, con lo que el modelo propuesto por Gairín, Muñoz y Oller (2012) se muestra pertinente. El mismo estudio identifica un conjunto de factores que alteran las calificaciones propuestas por los correctores y que podrían incluirse en el modelo original para refinarlo y obtener una variabilidad menor en las calificaciones finales obtenidas.

En este artículo presentamos una revisión de ese estudio, basado en la reelaboración de la guía de corrección utilizada considerando los factores distorsionadores que provocan una mayor variabilidad en las calificaciones obtenidas. Esta nueva guía se valida a partir de un estudio empírico en el que analizamos la reducción de la variabilidad conseguida en las calificaciones.

Evolución del modelo de Gairín Muñoz y Oller

Gairín, Muñoz y Oller (2012) propusieron una aportación teórica para la calificación de exámenes de matemáticas a partir de unos fenómenos que detectaron en la corrección de exámenes (modelo GMO en adelante). Esta aportación da lugar a unos criterios de corrección del tipo descuento por error (Watts y García, 1999). Sin embargo, se diferencia de este tipo de corrección en el hecho de que no concreta de forma exacta la cantidad de puntos a restar por cada error, sino que delimita cual es la puntuación máxima que el corrector puede quitar por un conjunto de errores; además, esta puntuación varía según la categoría en la que estén dichos errores tal y como muestra la siguiente figura:

Figura 1. Aportación teórica de Gairín, Muñoz y Oller (2012)



Además de establecer hasta que punto deberían penalizarse los errores según la tarea en la que se dan, estos autores observaron que se dan dos tipos de preguntas en los exámenes de matemáticas de las PAU que denominan preguntas de categoría 1 (C1) y preguntas de categoría 2 (C2). Las preguntas C1 son aquellas cuyo objetivo principal es el dominio de una técnica de cálculo; se necesita realizar tareas principales y auxiliares generales para poder resolverla. En las preguntas de categoría C2 el objetivo principal es el dominio de un concepto matemático; se necesita realizar tareas principales, auxiliares específicas y auxiliares generales para resolver este tipo de pregunta.

En Mengual, Gorgorió y Albarracín (2013) se presenta una prueba empírica de una primera concreción de unos criterios redactados a partir del modelo GMO. En su trabajo, cuatro correctores calificaron las respuestas que 67 alumnos dieron a las preguntas A3 y A4 de la prueba de matemáticas de las PAU de Zaragoza de Septiembre del 2010. Estudiando dichas correcciones, estos autores observaron que al aplicar los criterios propuestos se reducía la variabilidad entre las calificaciones de los diferentes correctores de un mismo examen de matemáticas.

En particular, en las preguntas de tipo C1 la variabilidad disminuyó más que en las preguntas C2, pero en ambas se documenta una mejora considerable en comparación con la variabilidad sin los criterios de corrección redactados a partir del modelo GMO. El porcentaje de respuestas para las que las calificaciones corregidas coincidían completamente oscila entre el 40% y el 70% y se constata que un 91.5% de las respuestas de los alumnos son calificadas con una desviación típica menor al 15%. A pesar de esta mejora, Mengual, Gorgorió y Albarracín (2013) detectaron unos fenómenos que introducen variabilidad en las calificaciones de las respuestas de los alumnos. Estos fenómenos aparecen cuando el alumno:

F1. Interrumpe la respuesta de una pregunta de clase C2 en cualquier punto del proceso de resolución.

F2. Propone una mala justificación de una solución o presenta una mala expresión final de ésta.

F3. Presenta resultados de cálculos no justificados.

F4. Manifiesta errores conceptuales que no afectan a la resolución pero que muestran un determinado desconocimiento.

■ Objetivos del Estudio

A partir de la evolución del modelo GMO y con el propósito de contribuir al desarrollo de una herramienta que permita valorar al alumnado de forma más uniforme y, por tanto a una evaluación más justa, los objetivos del estudio que presentamos son:

- Desarrollar un refinamiento del modelo de Gairín, Muñoz y Oller (2012) para la calificación de exámenes de matemáticas a partir de los fenómenos detectados en el estudio de Mengual, Gorgorió y Albarracín (2013).
- Contrastar si se consigue disminuir la variabilidad en las correcciones de los casos que presentan mayores dificultades de calificación a los correctores.

■ Metodología

Para llevar a término el estudio desarrollamos una investigación mixta, puesto que pretende analizar descriptivamente datos cuantitativos para determinar donde se produce variabilidad en las correcciones para después interpretar cualitativamente qué fenómenos conducen a dicha variabilidad (Godino, et al.,

2011). Para lo cual se estudian las calificaciones que otorgan 9 profesores a los exámenes de las PAU que generaron mayor variabilidad. También se tienen en cuenta las calificaciones otorgadas por el corrector oficial que corrigió dichas pruebas en esa convocatoria.

A partir del estudio llevado a cabo por Mengual, Gorgorió y Albarracín (2013) se eligieron las 15 de las 67 respuestas a las preguntas A3 y A4 de la prueba de matemáticas de las PAU de Zaragoza de Septiembre del 2010 que generaron mayor variabilidad por apartados y en la nota final de la prueba. Para el resto de respuestas, que generaron una variabilidad muy reducida, se considera que la concreción de los criterios utilizados por Mengual, Gorgorió y Albarracín (2013) es satisfactoria.

Para guiar la calificación, se rediseña una nueva concreción del modelo de calificación de Gairín, Muñoz y Oller (2012) considerando los fenómenos detectados en Mengual, Gorgorió y Albarracín (2013). Participan 9 correctores a los que se les pide que califiquen las respuestas de los alumnos dos veces: la primera sin especificar unos criterios de corrección más allá de los propios de la prueba original (SC) y la segunda utilizando la concreción rediseñada (GMO2).

■ Nueva concreción de criterios de corrección

En base a los fenómenos detectados en Mengual, Gorgorió y Albarracín (2013), se refinan los criterios de corrección atendiendo también al formato y presentación. En primer lugar, se decide crear una primera página introductoria a este tipo de evaluación dónde se explica a los correctores la aportación teórica GMO y dónde se incluyen los fenómenos detectados.

Para el primer fenómeno detectado –el alumno interrumpe la respuesta en cualquier punto del proceso de resolución– se especifica en la introducción que el corrector debe decidir qué fracción de la puntuación máxima de la actividad da en base a la relevancia de las tareas que ha contestado. En relación con los fenómenos 2, 3 y 4 –el alumno hace una mala justificación de la solución, o presenta una mala expresión final de ésta; presenta unos cálculos no justificados; o comete errores conceptuales que no afectan a la resolución pero que demuestran desconocimiento– el corrector debe tener en cuenta que para poder valorar una tarea, ésta debe estar debidamente justificada. Por ejemplo, los resultados de cálculos que no aparecen en la respuesta de los alumnos, no deberían tenerse en cuenta. Asimismo, si la respuesta de los alumnos contiene tareas que no conducen a la solución no se deben valorar.

A continuación y antes de definir los criterios de evaluación, la guía presenta unos cuadros que pretenden reflejar el camino que alumno podría seguir para dar respuesta a cada pregunta, con el fin de que este recurso pueda orientar al corrector. Además de tener en cuenta estas consideraciones, se optó por un formato de presentación de los criterios para que resultasen sencillos a los correctores. Este formato también se explica en la página introductoria de los criterios. A continuación, en la tabla1, se muestra cómo son los criterios de evaluación para el apartado A4 a) atendiendo al nuevo formato de presentación:

A4. a) Calcular el plano determinado por los puntos (1,0,0), (0,1,0) y (0,0,1). (*1 punto*)

Tabla 1.

Se puede penalizar con 1 punto por errores en: - Concepto: 3 puntos determinan un plano. - Procedimentales: Procedimiento para el cálculo del plano definido por tres puntos.	→ Nos indica que podemos finalizar con el proceso de evaluación penalizando con la totalidad de los puntos si el alumno comete errores en las tareas principales recogidas debajo de esta casilla.
Quitar como mucho hasta 0,7 puntos por el CONJUNTO de errores en: - Cálculo de dos vectores independientes con los 3 puntos - Utilizar un punto y dos vectores para dar la ecuación del plano	→ Nos indica que podemos penalizar hasta con 2/3 de los puntos por el conjunto de errores en las tareas auxiliares específicas recogidas debajo de esta casilla.
Quitar como mucho hasta 0,3 puntos por el CONJUNTO de errores en: - De tipo aritmético	→ Nos indica que podemos penalizar hasta con 1/3 de los puntos por el conjunto de errores en las tareas auxiliares generales recogidas debajo de esta casilla.

Análisis

En la Tabla 1 se muestran los resultados del porcentaje de casos (respuestas de alumnos) en los que la desviación típica de las calificaciones que dan los correctores disminuye al utilizar los criterios GMO2 respecto a la corrección sin criterios específicos (SC) por apartados. También se incluye el porcentaje de casos en los que la calificación en media prueba (A3 y A4 suponen la mitad de los puntos de una examen).

Tabla 2.

	A3		A4			*Nota en media prueba*
	a)	*b)*	*a)*	*b)*	*c)*	
Porcentaje	73,3 %	66,6 %	80%	73,3 %	66,6 %	66,6 %

Los resultados obtenidos en el estudio permiten afirmar que se produce una reducción de la variabilidad en las calificaciones de la prueba de matemáticas de las PAU al usar los criterios GMO2 respecto a la corrección con SC disminuye desde un punto de vista global en un 66,6% de los casos.

A continuación se analizan aquellas respuestas de los alumnos para las que los correctores asignan puntuaciones divergentes de las que mostramos un caso. En una pregunta de cálculo de extremos de funciones, un alumno comete un error en una tarea auxiliar general, que no debería suponer más de un tercio de penalización sobre 1.5 puntos posibles. Las calificaciones obtenidas utilizando los criterios GMO2 son las que se muestran en la Tabla 2.

Tabla 2. Una muestra de puntuaciones con criterios GMO2

C1	*C2*	*C3*	*C4*	*C5*	*C6*	*C7*	*C8*	*C9*
1	1,25	0,5	0	0,75	0,5	1,25	0,3	1

Como se puede observar, la mayoría de los correctores penalizan en exceso el error sin seguir fielmente los criterios especificados.

El análisis de los casos en los cuales se ha producido una variabilidad mayor con los criterios GMO2 que sin criterios específicos (SC) o se ha mantenido una variabilidad significativa, observamos que en varias ocasiones la diferencia la marcaba un solo corrector, poniendo una nota que difería significativamente del resto de calificaciones. De esta forma, no podemos asegurar que los correctores seguirán unos criterios establecidos, pues hemos visto que en casos aislados la mayoría de los correctores no aplicaban correctamente los criterios GMO2, y puntuaban en exceso o defecto las respuestas de los alumnos.

Por todo lo expuesto concluimos que los criterios GMO2 presentan una concreción que se muestra efectiva para minimizar la variabilidad en las calificaciones de una prueba de matemáticas. Al mismo tiempo, el análisis de los casos conflictivos muestra que la introducción de criterios por ella misma presenta limitaciones para reducir completamente la variabilidad, por lo que consideramos que la formación del profesorado debería ser el factor clave para conseguir pruebas con calificaciones más consistentes.

Conclusiones

A partir de los resultados obtenidos en el estudio podemos afirmar que la variabilidad en las calificaciones de la prueba de matemáticas de las Pruebas de Acceso a la Universidad disminuye en un porcentaje elevado de los casos estudiados. En concreto, se consigue mejorar la variabilidad en la suma global total de todos los apartados. El hecho de poder controlar la variabilidad en cada apartado permite que al final de la prueba la variabilidad entre los correctores haya mejorado en un 66,6% de los casos.

Tal y como afirman Gairín, Muñoz y Oller (2012) al poner estos límites el corrector debe dejar a un lado sus creencias y no penalizar más de lo que los criterios GMO explican. Por tanto, unos criterios de corrección más precisos elaborados a partir de la aportación teórica de Gairín, Muñoz y Oller (2012) mejoran la variabilidad en la corrección de la prueba de matemáticas, tal y cómo sugerían Cuxart, et al. (1997). Ahora bien, si tenemos en cuenta la suma de todas las desviaciones típicas, la variabilidad mejora en un 80% de los casos.

Si nos fijamos en los casos en los cuales se ha producido una variabilidad mayor con los criterios GMO2 que sin criterios o se ha mantenido una variabilidad significativa, observamos que en varias ocasiones la diferencia la marcaba un solo corrector, poniendo una nota significativamente más elevada que el resto. Asimismo, no podemos asegurar que los correctores seguirán unos criterios establecidos, pues hemos visto que en casos aislados la mayoría de los correctores no aplicaban correctamente los criterios GMO2, y puntuaban en exceso o defecto las respuestas de los alumnos. Además, en los casos en los que se

producía variabilidad con los criterios GMO2, ésta estaba asociada a alguno de los cuatro fenómenos detectados en Mengual, Gorgorió y Albarracín (2013) y los correctores no se ajustaban a las especificaciones propuestas en la introducción de los criterios de evaluación, este hecho puede darse debido a la escasa familiaridad que los correctores tienen con este tipo de criterios de corrección.

Los profesores participantes en el estudio no han tenido formación específica sobre cómo evaluar exámenes de matemáticas y aplican los criterios GMO2 por primera vez para este estudio. Por ello consideramos que los errores de corrección detectados podrían disminuir si los correctores recibieran una formación específica para su uso y los utilizaran de forma habitual.

De esta forma, la concreción de los criterios presentada en este trabajo proporciona una herramienta de calificación que ha mostrado empíricamente un alto nivel de efectividad en su propósito y invitamos a la comunidad educativa a considerar su inclusión en las guías de corrección de las futuras pruebas de evaluación de contenidos matemáticos.

Agradecimiento: Este trabajo está bajo el amparo del proyecto EDU2013-4683-R. Los autores forman parte del Grupo de Investigación Educació Matemàtica i Context: Competència Matemàtica (EMiC:CoM), con referencia SGR2014-723 (Generalitat de Catalunya).

Referencias bibliográficas

Casanova, M. A. (1998). Evaluación, concepto, tipología y objetivos. En M.A. Casanova. *La evaluación educativa* (pp.67-102). México: SEP-Muralla. Disponible en http://www.reformasecundaria.sep.gob.mx/espanol/pdf/ evaluacion/casanova/casanova3.pdf

Cuxart, A., Martí-Recober, M. y 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. *Revista de Educación*, *314*, 63-88.

Gairín, J.M., Muñoz, J.M., y Oller, A.M. (2012). Propuesta de un modelo para la calificación de exámenes de matemáticas. En A. Estepa, A. Contreras, J. Deulofeu, M. C. Penalva, F. J. García y L. Ordóñez (Eds.), *Investigación en Educación Matemática XVI* (pp. 261-274). Baeza: SEIEM

Godino, J. D., Carrillo, J., Castro, W. F., Lacasta, E., Muñoz-Catalán, M. C., Wilhelmi, M.R. (2011). Métodos de investigación en educación matemática: Análisis de los trabajos publicados en los simposios de la SEIEM (1997-2010). En M. Marín et al (Eds.), *Investigación en Educación Matemática XI* (pp.33-46). La Laguna: SEIEM.

Mengual, E., Gorgorió, N. y Albarracín, L. (2013). Validación de un instrumento para la calificación de exámenes de matemáticas. En A. Berciano, G. Gutiérrez, A. Estepa y N. Climent (Eds.), *Investigación en Educación Matemática XVII* (pp. 367-381). Bilbao: SEIEM.

Watts, F., y García, A. (1999). Control de calidad en la calificación de la prueba de inglés de selectividad. *Aula abierta*, *73*, 173-190.