DISEÑO DE CRITERIOS PARA REDUCIR LA VARIABILIDAD EN LA CALIFICACIÓN DE EXÁMENES DE MATEMÁTICAS EN PRUEBAS DE ACCESO A LA UNIVERSIDAD

Elena Mengual, Lluís Albarracín, José María Muñoz-Escolano, Antonio M. Oller-Marcén y Núria Gorgorió

Este artículo presenta el proceso de construcción de criterios para calificar exámenes de matemáticas con el propósito de reducir la variabilidad en las calificaciones otorgadas por distintos correctores. Desde la identificación de fenómenos no deseados en el proceso de corrección, se proporcionan directrices teóricas para generar criterios de calificación que se prueban empíricamente y se refinan en dos estudios consecutivos. Los resultados del estudio muestran que los Criterios de Calificación por Categorización de Tareas propuestos disminuyen de forma efectiva la variabilidad en las calificaciones en pruebas de matemáticas.

Términos clave: Calificación; Criterios de calificación; Exámenes de matemáticas; Fiabilidad entre correctores; Pruebas de acceso a la universidad

Criteria design to reduce variability in the grading of mathematics exams in tests of university entrance

This article presents the process of construction of criteria to qualify math tests in order to reduce the variability in the grades given by different correctors. From the identification of undesirable-phenomena in correction process, we provide theoretical guidelines to generate qualification criteria that are empirically tested and refined in two consecutive studies. The results of the study show that Rating Criteria by Categorization of proposed tasks effectively decrease the variability in the scores in math tests.

Keywords: Corrector's reliability; Entrance tests to University; Math exams; Qualification; Qualification criteria

Mengual, E., Albarracín, L., Muñoz-Escolano, J. M., Oller-Marcén, A. M. y Gorgorió, N. (2019). Diseño de criterios para reducir la variabilidad en la calificación de exámenes de matemáticas en pruebas de acceso a la universidad. *PNA 13*(2), 62-83.

En múltiples países el acceso a la universidad se condiciona a la superación de exámenes de acceso, particulares para cada facultad como en China, India, Rusia, Brasil o Estados Unidos, o bien pruebas de acceso generalizado como en Francia o Italia (Carnoy, Loyalka, Dobryakova, Dossani, Kuhns y Wang., 2013). Por otro lado, el modo en que cada país organiza las pruebas, gestiona el proceso de calificación y comunica los resultados de las mismas a los actores implicados, es muy diverso (Klein y van Ackeren, 2011).

En España, a mediados del siglo pasado se implantaron las pruebas de acceso a la universidad, que se caracterizan por establecer un mecanismo de selección y priorización de los estudiantes que pretenden acceder a estudios universitarios. Estas pruebas constan de un conjunto de exámenes que miden los aprendizajes de los alumnos en la etapa preuniversitaria y entre ellas se encuentran exámenes de matemáticas. Las pruebas de acceso han generado una preocupación social patente desde hace años, especialmente porque el acceso a una plaza de los estudios deseados a veces se consigue por tener unas centésimas más que otro aspirante (Escudero y Bueno, 1994). Por lo tanto, sería deseable que las calificaciones obtenidas por los estudiantes fueran consistentes con su nivel de aprendizaje (Gaviria, 2005). Sin embargo, diferentes estudios han señalado la existencia de diferencias entre correctores, que son los encargados de calificar cada prueba, y la necesidad de aumentar la fiabilidad de las mismas mejorando el sistema de evaluación y calificación (Escudero y Bueno, 1994; Grau, Cuxart y Martí-Recober, 2002).

Los exámenes de matemáticas de las Pruebas de Acceso a la Universidad en España (PAU) son un ejemplo de prueba de evaluación externa con gran influencia en la enseñanza y aprendizaje de las matemáticas en el Bachillerato (Contreras, Ordóñez y Wilhelmi, 2010), habiéndose comprobado que ejercen una influencia directa sobre los currículos implementados en las aulas y las metodologías docentes usadas (Rodríguez-Muñiz, Díaz, Mier y Alonso, 2016). Sobre la forma en la que se determinan las calificaciones de los exámenes de matemáticas de estas pruebas, se conocen algunos estudios estadísticos (Escudero y Bueno, 1994; Nortes, Nortes y Lozano, 2015) que parecen señalar que no hay diferencias significativas entre tribunales. Sin embargo, trabajos como el de Grau et al. (2002) o los de Gairín, Muñoz y Oller (2012, 2013) sí parecen mostrar la influencia del corrector en el proceso de calificación de exámenes. Desde un punto de vista cualitativo, no existen demasiados trabajos que estudien cómo desempeñan los profesores la tarea de calificar los exámenes de matemáticas en las pruebas de acceso. Se puede tomar como referencia el estudio de Bergeron (2015), en el que se comparan los contenidos, la estructura de las preguntas del examen y el tipo de rúbricas de evaluación de pruebas de acceso a la universidad en diversos países, pero sin estudiar la acción de los correctores en el proceso de calificación.

Es necesario tomar en cuenta que el trabajo de corrección de exámenes habitualmente no está incluido en los planes de formación de los docentes.

Además, en el caso concreto de las pruebas de matemáticas de las PAU, la corrección y calificación no son tareas triviales. En particular, se ha constatado que los correctores no corrigen igual la primera que la última prueba (Casanova, 1998). Además, algunos trabajos constatan la disparidad en los criterios proporcionados a los correctores dependiendo del distrito universitario (Boal, Bueno, Leris y Sein-Echaluce, 2008). Este contexto motiva nuestra investigación, orientada a la construcción y refinamiento de criterios que permitan controlar la variabilidad de las calificaciones de pruebas de matemáticas en pruebas externas o similares.

EVALUACIÓN EN PRUEBAS ESCRITAS

La evaluación constituye la herramienta principal para comprobar si se ha producido aprendizaje por parte de los alumnos y debe permitir conocer el grado del domino alcanzado por estos atendiendo a los objetivos propuestos, y determinar si el proceso de enseñanza ha sido adecuado para alcanzar dichos objetivos (Cantón y Pino-Juste, 2011). El concepto de evaluación se utiliza en numerosos contextos dentro y fuera del mundo educativo, por lo que es necesario especificar qué entendemos por evaluación en el trabajo y a qué tipo específico de evaluación nos referimos.

En nuestro estudio nos centraremos en la evaluación a partir de la calificación en una prueba (Castillo y Cabrerizo, 2003) puesto que, al evaluar la prueba de matemáticas de las PAU, lo que estamos haciendo es valorar el aprendizaje de cada alumno y en qué grado ha alcanzado los conocimientos, destrezas o habilidades matemáticas que se miden en esta prueba a partir de las evidencias que quedan sobre el papel. Pese a la diversidad de enfoques y de instrumentos existentes para la evaluación y pese a que calificar exámenes de respuesta abierta es un proceso complejo (Wang y Cai, 2018), lo cierto es que, en la práctica, es muy común el uso de este tipo de pruebas para la evaluación de los alumnos (Cárdenas, Blanco, Guerrero y Caballero, 2016; Moro-Egido, 2016). La evaluación tiende así a identificarse con la calificación de dichas pruebas (Senk, Beckmann y Thompson, 1997). Además, Duncan y Noonan (2007) señalan la especificidad del área de Matemáticas a este respecto.

En el caso de las pruebas de acceso a la universidad se da la circunstancia de que los correctores de la prueba no conocen a los alumnos que están calificando. Esto atenúa algunos posibles sesgos, como el llamado efecto "halo" (Rassmussen, 2008). Sin embargo, surgen otros aspectos importantes que deben ser tenidos en cuenta, siendo la fiabilidad entre correctores uno de los principales.

La fiabilidad entre correctores (Lane, Stone, Ankenmann y Liu, 1994) se entiende como el grado de coincidencia de las calificaciones otorgadas por distintos correctores a iguales producciones de un alumno. Existen diversos

estudios que constatan las discrepancias que surgen cuando diversos grupos de correctores actúan sobre el mismo conjunto de exámenes (Arnal-Bailera, Muñoz-Escolano y Oller-Marcén, 2016; Fitzpatrick, Ercikan, Yen y Ferrara, 1998; Meier, Rich y Cady, 2006). Este hecho ilustra la necesidad de identificar factores que puedan producir dichas discrepancias para, de ese modo, poder atenuar su influencia y mejorar la fiabilidad de las calificaciones. Wang y Cai (2018) señalan hasta cuatro factores que influyen en las correcciones: la experiencia docente del corrector, el nivel educativo en que posee dicha experiencia, la naturaleza de las respuestas de los alumnos (apareciendo las mayores diferencias en aquellas que contienen errores matemáticos) y las creencias sobre la enseñanza y el aprendizaje de las matemáticas.

Para situar nuestra investigación dentro de los diferentes tipos de evaluación tenemos en cuenta la clasificación propuesta por Casanova (1998). De esta forma el examen de matemáticas es parte de la prueba de acceso a la universidad, con lo que se trata de una evaluación sumativa y final. Si nos centramos en la evaluación según su normotipo, es decir, el referente que tomamos para evaluar al alumno, estamos ante una evaluación nomotética criterial. Se trata de una evaluación nomotética porque la valoración del sujeto se realiza sin conocer las capacidades del alumno y sus posibilidades de desarrollo. Dentro de la evaluación nomotética nos centramos en un tipo de evaluación criterial, puesto que se pretende evaluar al alumno en base a unos criterios externos bien marcados, concretos y claros. Estos criterios permiten valorar al alumnado de forma homogénea, y por ende, determinar en qué grado se han logrado los conocimientos, destrezas o habilidades matemáticas.

Para Sanmartí (2007) un criterio de evaluación es una norma que tomamos como referencia para interpretar la información recogida en la evaluación, es decir, para analizarla y emitir un juicio. Por otro lado, para Giménez (1997) un criterio es un "método para realizar evaluaciones y ajustarlas a un sistema de categorías establecido" (p. 27), es decir, son las afirmaciones que precisan el grado y tipo de aprendizaje que se ha alcanzado. Desde un punto de vista métrico, los criterios definidos para la evaluación han de ser objetivos, claros, comprensibles, preferiblemente cuantitativos, fiables y válidos (Muñiz y Fonseca-Pedrero, 2009). Para que unos criterios de corrección sean útiles deben manifestar un contenido de carácter general y una forma de juzgar las respuestas de los estudiantes (Giménez, 1997).

Estructura y objetivos del estudio

Este artículo presenta una investigación desarrollada en diversas fases, desde la detección de factores que introducen variabilidad en la calificación de pruebas a la construcción y el proceso de refinamiento y validación de unos criterios de corrección. A continuación, describimos la estructura de la investigación

presentada en este artículo y los objetivos de cada una de las tres fases desarrolladas.

La fase I se centra en la detección de los fenómenos que introducen variabilidad en las calificaciones que otorgan los correctores de los exámenes de matemáticas de las pruebas de acceso. A partir del análisis de la naturaleza de estos fenómenos se presenta una categorización de las tareas matemáticas a evaluar en estas pruebas y una propuesta teórica de modelo de calificación basada en la penalización de errores.

En la fase II concretamos el modelo de calificación propuesto en un conjunto de criterios que denominamos Criterios de Calificación por Categorización de Tareas (CCCT1) y realizamos un primer estudio empírico para contrastar si la variabilidad en las calificaciones disminuye con el uso de estos criterios. A partir del análisis de los datos se observa una importante disminución de la variabilidad, pero se detectan fenómenos en las respuestas de los alumnos a la prueba que no permiten que la disminución de la variabilidad sea óptima.

En la fase III se consideran los fenómenos detectados en la fase II y se reelaboran los criterios de calificación anteriores generando una nueva concreción de estos que denominamos CCCT2. Esta nueva concreción de los criterios es utilizada en un segundo estudio en el que se comparan las calificaciones que dan los correctores utilizando los criterios definidos (CCCT2) con las calificaciones cuando no se les proporcionan más criterios que los que se daban en la prueba original (CO).

El objetivo general de este trabajo es desarrollar unos criterios de calificación específicos para pruebas escritas de matemáticas que permitan reducir la variabilidad de calificaciones entre correctores. Para ello nos fijamos tres objetivos específicos para cada una de las fases de la investigación especificadas en el apartado anterior y que son los que se detallan a continuación.

- Identificar fenómenos no deseables en el proceso de calificación de exámenes de matemáticas y elaborar un modelo de calificación diseñado para evitarlos.
- Contrastar si el uso de los criterios CCCT1 reduce la variabilidad de las calificaciones que asignan los correctores e identificar los fenómenos que contribuyen a que se mantenga la variabilidad en las calificaciones.
- Desarrollar los criterios CCCT2 a partir de los fenómenos detectados en la fase II y contrastar si el uso de CCCT2 consigue disminuir la variabilidad en las calificaciones de los casos que presentan mayores dificultades de calificación a los correctores.

Fase I. Detección de fenómenos

La problemática que genera la necesidad de este estudio surge del análisis de los resultados de dos tribunales de Matemáticas II en junio de 2010 en la que se detecta una diferencia de 1,88 puntos sobre 10 en la media de las calificaciones

otorgadas por dos correctores distintos sin ningún sesgo aparte de la actuación de cada corrector. Dado que los exámenes en los que se detectó esa diferencia habían sido asignados de forma aleatoria a los distintos correctores, y que estos tuvieron que aplicar los mismos criterios de corrección proporcionados por el armonizador de la prueba, cabe pensar que la diferencia detectada se debe en buena medida, a que existen cierta componente de subjetividad en la valoración de las respuestas de los alumnos derivada de algunos de los factores apuntados por Wang y Cai (2018).

A partir de esta situación se inició un estudio en el que se detectaron nueve fenómenos a partir de un análisis exploratorio de las calificaciones otorgadas por seis correctores en los tribunales de septiembre de 2010 a 409 exámenes de Matemáticas II y Matemáticas Aplicadas a las Ciencias Sociales en la Comunidad Autónoma de Aragón (Gairín et al., 2012; 2013). Los correctores de la prueba dispusieron de unos criterios de corrección que dejaban, en buena medida, a su interpretación la calificación a otorgar. A modo de ejemplo mostramos el enunciado de la pregunta A3a y su correspondiente criterio de calificación en la figura 1.

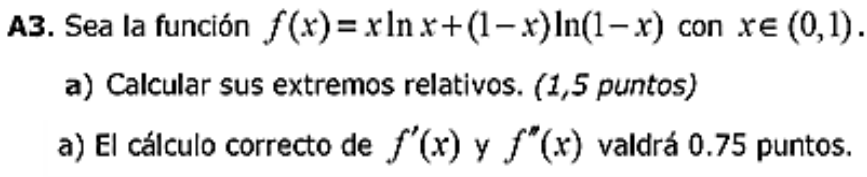

Figura 1. Enunciado y criterios de calificación de la pregunta A3a

Los fenómenos identificados en las actuaciones de los correctores (FC) se muestran en la tabla 1.

Tabla 1
Fenómenos identificados en las actuaciones de los correctores

Fenómeno	Descripción
FC1	Se califican erróneas respuestas correctas que utilizan conceptos distintos de los esperados.
FC2	Se califican erróneas respuestas correctas que utilizan técnicas de cálculo distintas de las esperadas.
FC3	Se consideran erróneas respuestas correctas que utilizan sistemas de representación distintos de los esperados.
FC4	La rigidez en el uso del lenguaje simbólico provoca que se califiquen como erróneas respuestas correctas.
FC5	Existen posiciones claramente diferenciadas entre distintos correctores al valorar la importancia de un mismo error.

Tabla 1
Fenómenos identificados en las actuaciones de los correctores

Fenómeno	Descripción
FC6	La importancia de los errores se establece con independencia del conocimiento matemático que se pretende evaluar.
FC7	La importancia de los errores se establece con independencia de las exigencias del enunciado de los problemas.
FC8	Al valorarse un error como muy grave, se da por finalizado el proceso de corrección.
FC9	Algunos correctores cometen errores al revisar los cálculos que figuran en las respuestas.

Nota. FC=Fenómeno en las actuaciones de los correctores.

Los cuatro primeros fenómenos están relacionados con las expectativas del corrector respecto a la respuesta correcta. Los fenómenos 5 a 8 afectan a cómo los correctores gestionan la presencia de un error en la respuesta de un alumno. El último fenómeno se corresponde con la aparición de "falsos positivos", es decir, casos en los que los correctores no detectan un error en la respuesta de un alumno. Algunos de estos fenómenos son difíciles de superar puesto que tienen que ver, en cierto modo, con las concepciones y creencias de los correctores (FC1-4) o con errores del propio corrector (FC9). No obstante, Gairín et al. (2013) proponen algunas sugerencias para superar estos cinco fenómenos. Los cuatro fenómenos restantes (FC5-8) ponen de manifiesto la necesidad de establecer claramente los objetivos que se pretenden evaluar en el problema, así como de enfatizar la importancia de la tarea donde aparece el error respecto del proceso global de resolución.

En este sentido, se propone una categorización de las tareas que intervienen en el proceso de resolución de un problema, así como un modelo de penalización de errores basado en dicha categorización (Gairín, Muñoz y Oller, 2012). En concreto, atendiendo a su cercanía con el objetivo principal del problema, se distinguen los siguientes tipos de tarea.

Tareas principales. Son aquellas que constituyen el objetivo principal de la calificación y pretenden valorar la comprensión del alumno sobre los contenidos matemáticos propios del curso en que se plantea el problema.

Tareas auxiliares. Son aquellas que aparecen en el proceso de resolución del problema, con la finalidad de obtener las informaciones necesarias para resolverlo. A su vez, distinguimos entre:

- Tareas auxiliares específicas. Están relacionadas con los contenidos matemáticos propios del curso en que se plantea el problema, pero

únicamente juegan un papel instrumental para alcanzar la solución del problema.

- Tareas auxiliares generales. Son aquellas que también juegan un papel instrumental en la solución del problema, pero están relacionadas con contenidos matemáticos recibidos por el alumno en su formación matemática anterior.

En función del tipo de tareas que están presentes en la resolución de una pregunta hablamos de dos categorías de preguntas (C1 y C2). Las preguntas C2 son más complejas desde un punto de vista conceptual y se caracterizan por la existencia de tareas auxiliares tanto específicas como generales. En cambio, las preguntas C1 no requieren la presencia de tareas auxiliares específicas y generalmente son aplicaciones directas de un procedimiento, por lo que la aplicación de dicho procedimiento es el objetivo principal de la calificación de la tarea.

A partir de esta categorización de tareas y la distinción de los tipos de preguntas, establecemos un marco que oriente el proceso de calificación de exámenes de matemáticas tratando de atenuar la presencia de los FC5-8 descritos anteriormente. En particular, proponemos el modelo de penalización de errores que se articula en base a las propuestas siguientes y que se esquematiza en la figura 2.

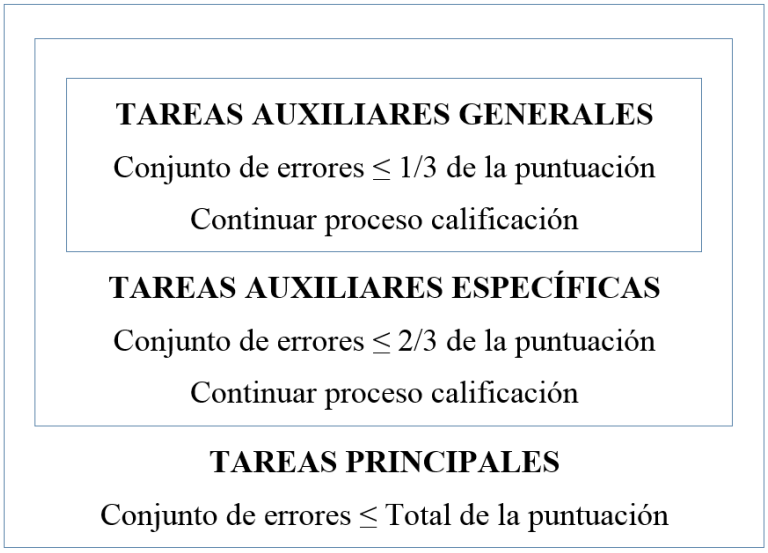

Figura 2. Esquema del modelo de criterios de calificación por categorización de tareas

El conjunto de todos los errores que comete un alumno al realizar tareas auxiliares generales no podrá penalizarse con un valor superior a un tercio de la puntuación asignada al problema.

Los errores cometidos en las tareas auxiliares específicas podrán penalizarse con un máximo de dos tercios de la calificación asignada a la respuesta correcta.

En este límite de penalización han de contabilizarse tanto los errores en las tareas auxiliares específicas como en las tareas auxiliares generales.

Los errores localizados en las tareas principales pueden penalizarse hasta con el cien por ciento de la puntuación total. En el caso de que en la resolución de un problema estén implicadas dos o más tareas principales, deberían asignarse puntuaciones independientes a cada una de ellas.

Este modelo, que denominamos Criterios de Calificación por Categorización de Tareas (CCCT), responde al tipo de criterio de descuento por error. En él no se fijan cuántos puntos restar exactamente (pues resulta muy complejo establecer penalizaciones concretas para cada tipo de error), sino que se establecen tres intervalos para marcar los valores entre los que ubicar las penalizaciones por los errores detectados.

Fase II. Modelo de calificación

A continuación, mostramos la primera parte del proceso para poner a prueba empíricamente el modelo que surge de la aportación teórica CCCT para la calificación de exámenes de matemáticas. Disponemos de las respuestas que dieron 76 alumnos a la prueba de matemáticas de las PAU de Zaragoza de septiembre del 2010. La prueba consta de dos opciones, A y B, entre las cuales el alumno debe decidir cuál realiza. Para la recogida de datos tomamos las pruebas correspondientes a la opción A, dado que fue escogida por 67 alumnos. Cada opción contiene 4 preguntas, cada una con varios apartados. Para nuestro estudio elegimos 2 de las 4 preguntas, concretamente las preguntas A3 y A4. Estas dos preguntas tienen diversos apartados de diferente naturaleza: A3a, A3b y A4a son preguntas de categoría C2 y A4b y A4c son preguntas de categoría C1, porque su resolución implica la mera aplicación de una fórmula (figura 3).

A3. Sea la función $f(x) = x\ln x + (1-x)\ln(1-x)$ con $x \in (0,1)$.

a) Calcular sus extremos relativos. *(1,5 puntos)*

b) Estudiar su crecimiento y decrecimiento y razonar si posee algún punto de inflexión. *(1 punto)*

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

b) Determinar el ángulo que forman los planos

$$\pi_1 \equiv \sqrt{2}x + y + z = 2 \quad \text{y} \quad \pi_2 \equiv z = 0 \qquad \textit{(0,75 puntos)}$$

c) Obtener el producto vectorial de $\vec{a} = (2,0,1)$ y $\vec{b} = (1,-1,3)$. *(0,75 puntos)*

Figura 3. Enunciados de las preguntas A3 y A4

Esta elección de preguntas nos permite elaborar unos criterios de calificación basados en el modelo teórico presentado para las dos categorías de pregunta consideradas.

Los datos recogidos son las calificaciones que otorgan 4 profesores que han intervenido como correctores en pruebas de matemáticas de acceso a la

universidad a las 67 respuestas estudiadas de las preguntas A3 y A4, utilizando la primera concreción de los criterios de calificación presentados. Mostramos esta concreción a continuación en forma de instrumento de recogida de datos del estudio general.

Instrumento para la recogida de datos de la fase II

Elaboramos una primera guía que concreta los criterios de corrección propuestos en el modelo de calificación para las preguntas A3 y A4, que constituyen la mitad de la calificación total de la prueba. A esta concreción de los criterios la llamamos CCCT1. A continuación, de los dos apartados que tiene la pregunta A3, mostramos el proceso de desarrollo de la guía para el primer apartado (ver enunciado en figura 1). Este apartado es de categoría C2, por tanto, para su resolución se necesita el dominio de un concepto matemático y el desarrollo de tareas principales, auxiliares específicas y auxiliares generales. El resto de criterios para los demás ejercicios se desarrollan de forma análoga.

El objetivo principal de A3a es evidenciar el dominio de un concepto matemático, en este caso de extremo relativo y su clasificación, es decir, determinar si el punto crítico es un máximo, un mínimo o un punto de inflexión. Antes de definir las tareas que requiere dar respuesta a esta pregunta, identificamos los pasos que el alumno podría seguir hasta la solución. Una posible secuencia, de las varias existentes, para resolver la pregunta queda recogida en la figura 4.

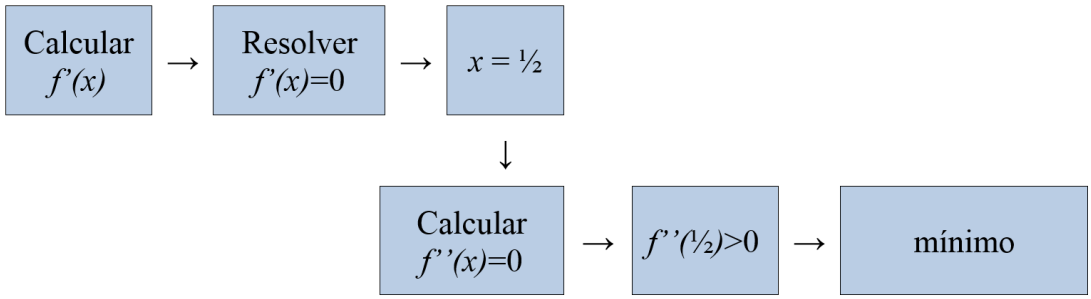

Figura 4. Diagrama de una posible secuencia de resolución de la pregunta A3a

Esta concreción muestra que existen dos procesos generales necesarios para resolver la pregunta: el primero es determinar el punto crítico y el segundo clasificarlo. En la tabla 2 se muestran los criterios de evaluación para este ejercicio.

Estos criterios pretenden acotar el campo de actuación del corrector y evitar penalizaciones excesivas por fallos en tareas matemáticas básicas. El resto de criterios correspondientes al resto de preguntas se desarrollaron de forma análoga y son los que se proporciona a los profesores participantes para que realicen las correcciones.

Tabla 2
Criterios de evaluación para la pregunta A3a

Tareas		Penalización
	Principales	
Conceptuales	Extremo relativo	Se puede penalizar con la totalidad por errores en tareas principales
	Clasificación de extremo relativo	
Procedimentales	Cálculo de extremos relativos	
	Auxiliares	
Específicas	Cálculo de derivadas	Como mucho 1 punto por el conjunto de errores en tareas auxiliares generales y específicas.
	Sustitución de un punto en una función	
Generales	Algebraicas: simplificar las derivadas, resolver ecuaciones y/o inecuaciones.	Como mucho 0,5 puntos de penalización por el total de los errores en tareas auxiliares generales.
	Aritméticas: realizar operaciones al sustituir el punto en la función.	

Análisis de datos y resultados de la fase II

Los cuatro profesores corrigieron las 67 respuestas de los alumnos utilizando la concreción de los criterios CCCT1. Los profesores reciben cada uno una copia de los mismos exámenes originales que no contienen marcas previas de corrección. De esta forma tenemos 4 calificaciones para las respuestas de los 67 alumnos en cada uno de los apartados de las preguntas A3 y A4. La tabla 3 muestra la media de las calificaciones que dan los correctores al total de preguntas calificadas.

Tabla 3
Calificaciones medias de cada corrector al conjunto de las preguntas A3 y A4

C1	C2	C3	C4
1,86	2,13	1,96	2,02

Nota. C=corrector.

Los valores recogidos en la tabla anterior muestran una diferencia máxima entre correctores de 0,27 puntos sobre un total de 5 puntos evaluados. Esto equivaldría, suponiendo que los correctores mantienen su comportamiento, a una diferencia

máxima de 0,54 puntos sobre 10, resultado que muestra que usando este modelo se han obtenido diferencias entre las puntuaciones menores que las detectadas sin su uso con anterioridad.

Aunque las medias de las calificaciones dadas por los diferentes correctores presenten una baja variabilidad, podemos observar casos concretos en los que la variabilidad se mantiene. Para estudiarlos se han calculado la media y la desviación típica de las calificaciones de los correctores a cada respuesta para poder observar aquellos casos en los que las calificaciones no se encuentran alrededor de un valor central —que representaría la calificación ideal— y presentan variabilidad en las calificaciones. La figura 5 muestra una parte de la hoja de cálculo que contiene esta información.

	A	B	C	D	E	F	G	H	I	J
1										
2		**EJERCICIO A3 a)**								
3										
4			Corrector 4	Coorector 3	Corrector 2	Corrector 1	Media (*M*)	Desviación típica (*DT*)		
5		100_003	1	1	1	1	1	0		Si nota < *M-DT*
6		100_004	1,5	1,4	1,5	0,75	1,2875	0,312999601		si nota > M+DT
7		100_006	0	0	0	0	0	0		
8		100_007	1,5	1,5	1,5	1,5	1,5	0		
9		100_008	0,75	0,75	0,75	0,75	0,75	0		
10		100_009	0	0	0	0	0	0		
11		100_010	1,5	1,5	1,5	1,5	1,5	0		
12		100_011	0	0	0	0	0	0		
13		100_012	1,5	1,2	1,25	1,5	1,3625	0,138631706		
14		100_013	1	1	1	1	1	0		
15		100_015	0	0	0	0	0	0		
16		100_016	1,5	1,5	1,5	1,5	1,5	0		
17		100_017	0	0	0,5	0	0,125	0,216506351		
18		100_018	0	0	0	0	0	0		
19		100_019					0	0		
20		100_020	0	0	0,25	0	0,0625	0,108253175		

Figura 5. Calificaciones de los correctores por apartados

Por ello, para cada uno de los apartados de las preguntas A3 y A4 se calcula la media y la desviación típica de las puntuaciones de los alumnos a partir de las correcciones que los cuatro profesores asignan a cada estudiante.

La variabilidad ideal entre las calificaciones es cero, pero entendemos que en el caso de calificar preguntas abiertas no va a ser posible conseguir este resultado. De esta forma, teniendo en cuenta que no existe en la literatura una discusión sobre la variabilidad máxima aceptada, tomamos como aceptable una desviación típica inferior al 10% de la puntuación total de cada apartado de A3 y A4. Con esta decisión marcamos un umbral en el que la calificación es consistente al asegurar que la desviación de la nota final del examen no exceda de 1 punto. Además, si consideramos las posibles compensaciones entre las calificaciones de los diversos apartados, la desviación total debería ser considerablemente menor. La figura 6 recoge la distribución de las calificaciones recogidas.

Las calificaciones de los correctores coinciden totalmente en un 59,6% de los casos estudiados y se puede observar que la desviación típica en el conjunto de las preguntas A3 y A4 es menor del 10% marcado como umbral en un 97% de los casos. Las respuestas de los alumnos con una desviación típica superior al 10% son consideradas de alta variabilidad y se estudian cualitativamente para detectar el origen de la variabilidad detectada. El análisis de dichas correcciones tiene como propósito estudiar si es posible detallar más los criterios de corrección para que disminuya la variabilidad.

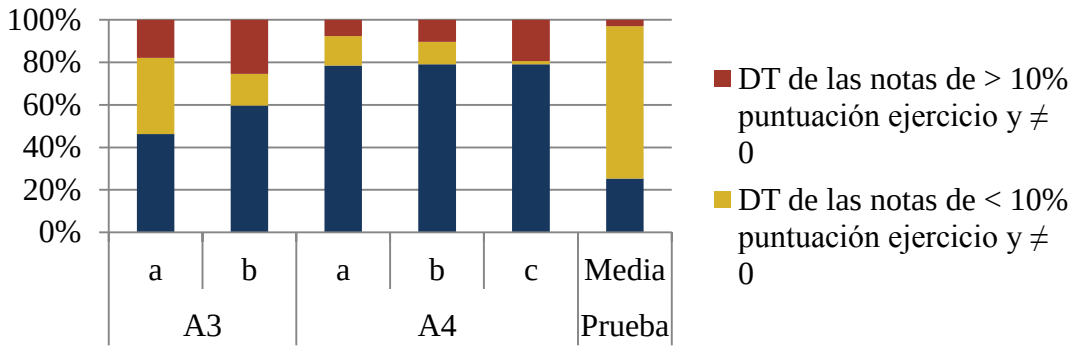

Figura 6. Distribución de la desviación típica de las calificaciones por apartados

El detalle del análisis cualitativo de las respuestas de los alumnos que generan variabilidad en las calificaciones recogidas nos permite identificar en estas respuestas cuatro fenómenos que afectan a la calificación (Mengual, Gorgorió y Albarracín, 2013). Los fenómenos identificados a partir de este análisis en las respuestas de los alumnos (FR) y que afectan a la variabilidad en las calificaciones utilizando los criterios CCCT1 son los siguientes:

- FR1: La respuesta de una pregunta de clase C2 se interrumpe en cualquier punto del proceso de resolución.
- FR2: La respuesta contiene una mala justificación de una solución o una mala expresión final de esta.
- FR3: La respuesta contiene resultados de cálculos no justificados.
- FR4: La respuesta contiene errores conceptuales que no afectan a la resolución pero que muestran un determinado desconocimiento.

Fase III. Criterios de calificación

A partir del estudio realizado en la fase II, el uso de los criterios CCCT1 muestra un potencial manifiesto para reducir la variabilidad en la mayoría de los casos a partir de utilizar unos criterios que respondan a los fenómenos FC5-8 identificados en la fase I. Las limitaciones en la reducción de la variabilidad de los CCCT1 ponen de manifiesto los fenómenos no deseados FR1-4 que generan la necesidad de refinarlos para crear una nueva concreción de criterios que denominamos CCCT2. Esta tercera fase del estudio se centra en la elaboración de unos criterios de calificación refinados y en su validación. Participan nueve

profesores expertos que corrigen las 15 pruebas que han presentado mayores valores de variabilidad en la fase II.

Instrumento para la recogida de datos de la fase III

Para la recogida de datos se refinan los criterios de corrección (CCCT2), a partir de incluir medidas para paliar los fenómenos detectados en la fase II. En primer lugar, se adjunta una primera página introductoria donde se explica a los correctores la aportación teórica presentada. En referencia al primer fenómeno detectado, en el cual el alumno interrumpe la respuesta de una pregunta en cualquier punto del proceso de resolución, se especifica en la introducción que el corrector debe decidir que 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, en los cuales el alumno hace una mala justificación de la solución, una mala expresión final de esta, unos cálculos no justificados o se demuestran errores conceptuales que no afectan a la resolución pero que demuestran un desconocimiento, se indica al corrector que, para poder valorar una tarea, esta 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, se muestran unos cuadros que pretenden reflejar los caminos que un alumno podría seguir para dar respuesta a cada pregunta, definiendo las tareas que componen la pregunta, con el fin de que este recurso pueda guiar al corrector. Además de tener en cuenta estas consideraciones, se optó por un formato de presentación de los criterios que resultasen más sencillos a los correctores. Este formato también se explica en la página introductoria de los criterios. En la figura 7, se muestra cómo son los criterios de evaluación para el apartado A3a atendiendo al nuevo formato de presentación.

Se puede penalizar con 1,5 puntos por errores en:
- Conceptuales: Extremo relativo - Procedimentales: Cálculo de extremos relativos.
Penalizar como máximo hasta 1 punto por el CONJUNTO de errores en:
- Cálculo de derivadas - Sustitución de un punto en una función
Penalizar como máximo hasta 0,5 puntos por el CONJUNTO de errores:
- De tipo aritmético - De tipo algebraico

Figura 7. Presentación de criterios de calificación de la pregunta A3a

Para la recogida de datos, los correctores realizaron en primer lugar una corrección utilizando los criterios oficiales que se proporcionaron para la prueba

en las PAU de Zaragoza de septiembre de 2010, que dejan una gran libertad de corrección al corrector. La segunda corrección se realizó utilizando la concreción CCCT2 de los criterios elaborados. De esta forma, los datos recogidos son dos puntuaciones de cada corrector para cada uno de los apartados de las preguntas A3 y A4 de la prueba de matemáticas de selectividad de Zaragoza de septiembre de 2010.

Análisis de datos y resultados de la fase III

Para cada respuesta de los alumnos y cada apartado se calculan la media y la desviación típica de las calificaciones dadas por los 9 correctores para compararlas con la variabilidad obtenida en la fase II. En la figura 8 se recoge el porcentaje de respuestas calificadas en esta fase en las que se disminuye la variabilidad en las correcciones como índice del éxito en la reducción de la variabilidad que producen los CCCT2 respecto a los CCCT1.

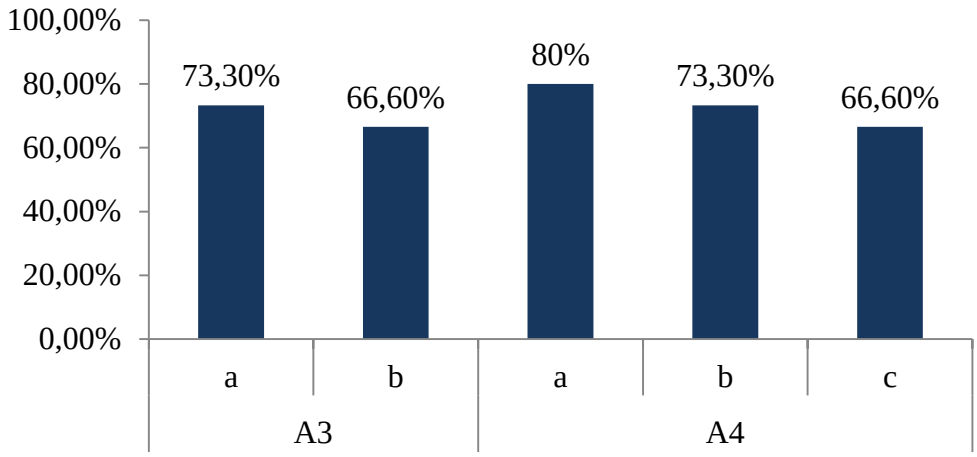

Figura 8. Porcentaje de respuestas en las que se disminuye la variabilidad con la corrección CCCT2

Se puede comprobar que las calificaciones generadas con los criterios CCCT2 reducen la variabilidad en más de un 66% de los casos estudiados, que son aquellos que representan mayores dificultades de calificación. De esta forma, los criterios CCCT2 muestran mayor consistencia que la corrección con los criterios originales de la prueba.

Al analizar los casos en los cuales la desviación típica con los criterios CCCT2 es mayor que con los criterios originales observamos que la variabilidad está asociada a una mala aplicación de los criterios CCCT2 por parte de los correctores, en especial en las preguntas de categoría C2. Por ejemplo, en el apartado A3a un alumno comete un error en una tarea auxiliar general y arrastra este error hasta el final, sin embargo, demuestra tener los conocimientos necesarios para calcular extremos relativos y clasificarlos. De acuerdo con los criterios CCCT2 los errores en las tareas auxiliares generales no se pueden

penalizar con más de una tercera parte de la puntuación total, en este ejercicio 0,5 puntos. La tabla 4 muestra las calificaciones para esta respuesta de correctores utilizando los criterios CCCT2. En ella se puede observar que penalizan en exceso el error detectado en esta resolución.

Tabla 4
Calificaciones CCCT2 a un caso conflictivo

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

Nota. C=corrector.

CRITERIOS DE CALIFICACIÓN DESARROLLADOS

A modo de producto del trabajo presentado en este artículo proponemos los CCCT2 como criterios para calificar pruebas de matemáticas en las que la reducción de la variabilidad de las calificaciones sea un aspecto relevante del proceso de evaluación.

Los criterios de evaluación CCCT2 propuestos indican hasta qué punto se puede penalizar un conjunto de errores y se basan en una jerarquización de las tareas que los alumnos podrían seguir para resolver cada pregunta de la prueba. Estas tareas pueden pertenecer a una de las siguientes dos categorías.

Tarea principal. Tareas que claramente constituyen el objetivo principal de la calificación.

Tareas auxiliares. Son de dos tipos:

- Específicas: aquellas tareas que juegan un papel instrumental para alcanzar la solución de un problema en el que aparecen tareas principales sobre contenidos específicos.
- Generales: Tareas matemáticas que ha realizado el alumno a lo largo de su formación matemática anterior.

Complementariamente se establecen dos consideraciones: (a) todas las actividades tienen una puntuación máxima. En el caso que un alumno deje inacabada una de las actividades, el corrector debe decidir qué fracción de esta puntuación máxima da en base a la relevancia de las tareas que ha contestado; y (b) para poder valorar una tarea debe estar debidamente justificada. Por ejemplo, resultados de cálculos que no aparecen en la respuesta de los alumnos, no deberían tenerse en cuenta. Si la respuesta de los alumnos contiene tareas que no conducen a la solución no se deben valorar. La figura 9 recoge el detalle genérico de estos criterios.

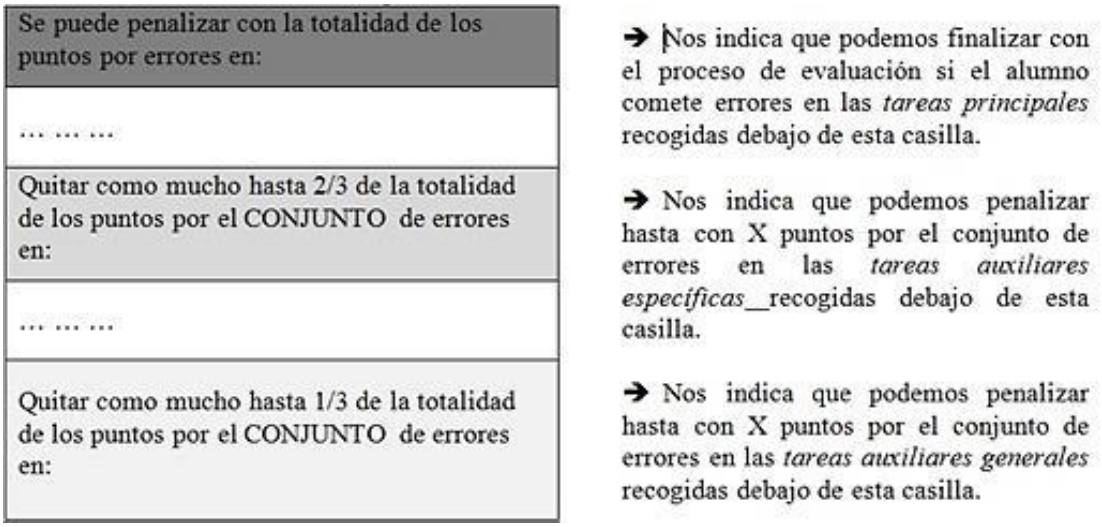

Figura 9. Detalle genérico de los criterios de corrección CCCT2

DISCUSIÓN Y CONCLUSIONES

En este estudio hemos presentado las diferentes fases que nos han permitido identificar fenómenos no deseados en la calificación de pruebas de matemáticas y la validación empírica del modelo de evaluación criterial basado en la categorización de tareas. El proceso de validación desarrollado muestra que, a parte del tipo de actividades propuestas y la actuación de los correctores, la naturaleza de las respuestas de los alumnos también es un aspecto generador de variabilidad en las calificaciones otorgadas, en línea con lo observado por Wang y Cai (2018).

Los resultados de este estudio parecen indicar que los criterios de calificación basados en el modelo CCCT pueden constituir una herramienta que reduzca de forma efectiva la variabilidad en las calificaciones que otorgan los evaluadores en pruebas de Matemáticas. Dado que las preguntas de las pruebas son abiertas y las respuestas de los alumnos presentan una gran diversidad, el proceso de calificación no es trivial para los correctores. Sin embargo, comprobamos que unos criterios de corrección más precisos reducen la variabilidad en la corrección de la prueba de matemáticas. Por lo tanto, consideramos un avance hacia la reducción de la variabilidad en las calificaciones el aporte presentado al contrastar su mayor nivel de fiabilidad respecto a los criterios utilizados en las pruebas reales.

Si atendemos los casos en los cuales se ha producido una variabilidad mayor con los criterios CCCT2 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 considerablemente más elevada que el resto. En los casos en los que se produce variabilidad con el uso de los criterios CCCT2, esta está asociada a alguno de los cuatro fenómenos detectados en la

fase II. Los profesores participantes en el estudio no han tenido formación específica sobre cómo evaluar exámenes de matemáticas y aplican los criterios CCCT2 por primera vez para este estudio. Por ello consideramos que los estos errores de corrección detectados podrían disminuir si los correctores recibieran una formación específica para su uso y los utilizaran de forma habitual. En la actualidad, se están desarrollando experiencias de formación de profesores en esta dirección (Arnal-Bailera, Cid, Muñoz-Escolano y Oller-Marcén, 2016).

Los CCCT2, dada la especificidad de su formulación, deberán ser elaborados para cada nueva prueba y requieren de un cambio de hábitos en los correctores, que hasta ahora gozaban de una mayor libertad a partir de unos criterios menos desarrollados y menos concretos. De hecho, los criterios CCCT2 pueden ser incorporados a la formación del profesorado de Educación Secundaria con el objetivo de complementar la formación de los futuros profesores en el ámbito de la evaluación.

Sin embargo, uno de los mayores problemas con el que nos enfrentamos en la búsqueda de criterios de calificación que permitan reducir la variabilidad es el hecho que en el caso de actividades matemáticas complejas estamos lejos de conseguir una reducción total de la variabilidad entre correctores. En este sentido los CCCT2 se muestran como una propuesta sólida que puede utilizarse para un amplio tipo de pruebas matemáticas, ya que toda actividad de evaluación podrá ser caracterizada siguiendo la categorización de tareas propuesta.

Este potencial de los CCCT2 abre la puerta a ser usado en las pruebas de matemáticas de la selectividad española pero también en otros entornos educativos. Por ejemplo, las pruebas de acceso a la universidad en Francia contienen actividades matemáticas más abiertas (en el sentido que existen diversas aproximaciones a su resolución) que las pruebas de matemáticas de la selectividad española. De la misma forma, algunas de las preguntas de las pruebas de matemáticas de la Evaluación del Bachillerato para el acceso a la Universidad (EBAU) son más abiertas que las preguntas de las PAU tradicionales, escenario en el que entendemos que los CCCT2 pueden aportar una guía de uso para controlar la variabilidad entre correctores, especialmente por el hecho que al tener que crear una nueva cultura de corrección, disponer de un tipo de criterios compartido puede ser beneficioso. Desde el punto de vista de la aplicación de los CCCT2, los criterios desarrollados para actividades de tipo C2 se deben poder ajustar a los estándares utilizados en pruebas escritas de matemáticas, aunque presenten una gran cantidad de aproximaciones a la resolución, ya que un corrector con la formación adecuada podrá establecer la caracterización de la resolución propuesta en cada caso y aplicarlos con un alto grado de fiabilidad.

Sobre las posibilidades de extender el uso de criterios CCCT2, entendemos que también pueden ser aplicados a pruebas de otras materias curriculares, como pueden ser Física, Química o Tecnología. La jerarquización de las tareas necesarias para resolver un problema o ejercicio de estas materias puede

realizarse de forma similar a la presentada en este estudio. En aquellas actividades con una fuerte carga matemática en su resolución, los conocimientos propios de la materia conformarán las tareas principales y las tareas auxiliares específicas de forma que los conocimientos matemáticos presentes en los cálculos actuarán como tareas auxiliares generales. Utilizando esta estructura para generar criterios se respeta la naturaleza instrumental de las matemáticas mientras las calificaciones se centran en evaluar los conocimientos propios de cada una de las materias.

Finalmente, entendemos que en este trabajo se presentan dos productos que responden a diferentes necesidades y que pueden tener recorrido en el ámbito de la investigación en educación. Por una parte, la jerarquización de tareas ya ha sido utilizada para desarrollar herramientas de análisis del contenido de libros de texto (Mengual, Gorgorió y Albarracín, 2017). Por otra parte, los criterios de calificación de pruebas CCCT2 deberían probarse en situaciones de evaluación real para validar su potencialidad en entornos reales y determinar si no se introducen nuevas problemáticas al proceso de calificación, como podría ser aumentar el tiempo necesario para que un corrector pueda proporcionar una calificación, que es determinante para que el uso de unos criterios de calificación sea aplicable en situaciones reales.

AGRADECIMIENTOS

La investigación que se presenta ha sido financiada por el proyecto EDU2017-82427-R (Ministerio de Economía, Industria y Competitividad, Spain). Los autores pertenencen a los grupos de investigación *Laboratori competència matemàtica en context (LABCOMeC)* (2017 SGR 497 - AGAUR, Generalitat de Catalunya) y *S36_17D-Investigación en Educación Matemática* financiado por el Gobierno de Aragón.

REFERENCIAS

Arnal-Bailera, A., Cid, E., Muñoz-Escolano, J. M. y Oller-Marcén, A. M. (2016). *Marking mathematics exams as a tool for secondary teacher training*. Ponencia presentada en 13th International Congress on Mathematical Education. Hamburgo, Alemania.

Arnal-Bailera, A., Muñoz-Escolano, J. M. y Oller-Marcén, A. M. (2016). Caracterización de las actuaciones de correctores al calificar pruebas escritas de matemáticas. *Revista de Educación, 371*, 35-60.

Bergeron, L. (2015). *IB Mathematics Comparability Study: Curriculum & Assessment Comparison*. Recuperado de: http://www.ibo.org/globalassets/publications/ib-research/dp/maths-comparison-summary-report.pdf

Boal, N., Bueno, C., Lerís, M. D. y Sein-Echaluce, M. L. (2008). Las habilidades matemáticas evaluadas en las pruebas de acceso a la universidad. Un estudio en varias universidades públicas españolas. *Revista de Investigación Educativa*, *26*(1), 11-23.

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

Cárdenas, J. A., Blanco, L. J., Guerrero, E. y Caballero, A. (2016). Manifestaciones de los profesores de matemáticas sobre sus prácticas de evaluación de la resolución de problemas. *Bolema*, *30*(55), 649-669. http://dx.doi.org/10.1590/1980-4415v30n55a17

Carnoy, M., Loyalka, P., Dobryakova, M., Dossani, R., Kuhns, K. y Wang, R. (2013). *University expansion in a changing global economy: Triumph of the BRICs?* Stanford, CA: Stanford University Press.

Casanova, M. A. (1998). *La evaluación educativa. Escuela básica.* Madrid, España: La Muralla.

Castillo, S. y Cabrerizo, J. (2003). *Evaluación educativa y promoción escolar*. Madrid, España: Pearson Education.

Contreras, A., Ordóñez, L. y Wilhemi, 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. https://doi.org/10.5565/rev/ec/v28n3.63

Duncan, C. R. y Noonan, B. (2007). Factors affecting teachers' grading and assessment practices. *Alberta Journal of Educational Research*, *53*(1), 1.

Escudero, T. y Bueno, C. (1994). Examen de selectividad: el estudio de un tribunal paralelo. *Revista de Educación*, *304*, 281-298.

Fitzpatrick, A. R., Ercikan, K., Yen, W. M. y Ferrara, S. (1998). The consistency between raters scoring in different test years. *Applied Measurement in Education*, *11*(2), 195-208. https://doi.org/10.1207/s15324818ame1102_5

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, Á. Contreras, J. Deulofeu, M. C. Penalva, F. J. García y L. Ordóñez (Eds.), *Investigación en Educación Matemática XVI* (pp. 261-274). Jaén, España: SEIEM.

Gairín, J. M., Muñoz, J. M. y Oller, A. M. (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.

Gaviria, J. L. (2005). La equiparación del expediente de Bachillerato en el proceso de selección de alumnos para el acceso a la universidad. *Revista de Educación, 337*, 351–387.

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

Grau, R., Cuxart, A. y 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.

Klein, E. D. y van Ackeren, I. (2011). Challenges and problems for research in the field of statewide exams. A stock taking of differing procedures and standardization levels. *Studies in Educational Evaluation*, *37*(4), 180-188.

Lane, S., Stone, C. A., Ankenmann, R. D. y Liu, M. (1994). Reliability and validity of a mathematics performance assessment. *International Journal of Educational Research, 21*(3), 247-266.

Meier, S. L., Rich, B. S. y Cady, J. (2006). Teachers' use of rubrics to score non-traditional tasks: factors related to discrepancies in scoring. *Assessment in Education*, *13*(01), 69-95. http://dx.doi.org/10.1080/09695940600563512

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

Mengual, E., Gorgorió, N. y Albarracín, L. (2017). Análisis de las actividades propuestas por un libro de texto: el caso de la medida. *REDIMAT*, *6*(2), 136-163.

Moro-Egido, A. I. (2016). La evaluación de las técnicas de evaluación en la enseñanza universitaria: la experiencia de macroeconomía. *Cultura y Educación*, *28*(4), 843-862.

Muñiz, J. y Fonseca-Pedrero, E. (2009). Construcción de instrumentos de medida para la evaluación universitaria. *Revista de Investigación en Educación*, *5*, 13-25.

Nortes, A. N., Nortes, R. y Lozano, F. (2015). Las correcciones en matemáticas en las pruebas de acceso a la universidad. *Educatio Siglo XXI*, *33*(3), 199-222.

Rasmussen, K. (2008). Halo Effect. En N. J. Salkind (Ed.), *Encyclopedia of Educational Psychology* (pp. 458-459). Thousand Oaks, CA: Sage. http://dx.doi.org/10.4135/9781412963848

Rodríguez-Muñiz, L. J., Díaz, P., Mier, V. y Alonso, P. (2016). Washback effect of university entrance exams in applied mathematics to social sciences. *PloS one*, *11*(12). https://doi.org/10.1371/journal.pone.0167544

Sanmartí, N. (2007). *10 ideas clave: evaluar para aprender* (Vol. 1). Barcelona, España: Graó.

Senk, S. L., Beckmann, C. E. y Thompson, D. R. (1997). Assessment and grading in high school mathematics classrooms. *Journal for Research in Mathematics Education*, *28*(2), 187-215.

Wang, N. y Cai, J. (2018). An investigation of how teachers score constructed-response mathematics assessment tasks. *Journal of Research in Education*, *28*(1), 1-29.

Elena Mengual
Universidad de Zaragoza
emengual@unizar.es

Lluís Albarracín
Universidad Autónoma de Barcelona
lluis.albarracin@uab.cat

José M. Muñoz-Escolano
Universidad de Zaragoza
jmescola@unizar.es

Antonio M. Oller-Marcén
Centro Universitario de la Defensa de Zaragoza
oller@unizar.es

Núria Gorgorió
Universidad Autónoma de Barcelona
nuria.gorgorio@uab.cat

Recibido: 15/07/ 2018. Aceptado: 24/11/2018.
doi: 10.30827/pna.v13i2.7740

ISSN: 1887-3987