FORMATOS DE EXAMEN Y OBJETIVIDAD EN LAS CALIFICACIONES ACADÉMICAS

ALFREDO FIERRO (*)
CARLOS FIERRO-HERNÁNDEZ (*)

Háblenme del tercer acto de Hamlet. El alumno no sabía y el profesor le dijo: Queda usted suspendido. Y la verdad es que también Shakespeare hubiera quedado suspendido porque la división en actos y escenas es posterior a él, una decisión de los editores. Y si a mí me pidieran que hablase de una página de El Aleph, también quedaría suspendido.

(Jorge Luis Borges)

RESUMEN. Se comparan los resultados de examen en modalidades distintas de evaluación y calificación de alumnos. Las modalidades estudiadas en dos grupos distintos de alumnos han sido: examen objetivo, a manera de test, de respuestas cerradas; corrección de frases; preguntas breves; tema de desarrollo. Excepto esta última modalidad, las demás han mostrado elevadas correlaciones entre sí. Respecto a ellas, no cabe decir que alguna sea significativamente más equitativa que otras en la calificación. Por otro lado, sin embargo, la elevada correlación entre las mismas es compatible con la circunstancia de que para alumnos concretos el resultado es muy distinto en una u otra modalidad.

En la universidad española los exámenes tienen un protagonismo académico, unas peculiaridades de formato y una trascendencia para la futura carrera profesional, que no encuentran parangón en otros sistemas universitarios. Una discusión sobre la validez de los formatos convencionales de examen, y de la consiguiente equidad de las calificaciones que derivan de ellos, difícilmente podría interesar a un profesor británico o germano, cuya tradición académica se adhiere a otros modos de evaluación de los conocimientos y destrezas de los estudiantes. Sin embargo, también en otros países, dentro o fuera del ámbito universitario, existen procedimientos selectivos que corresponden a un modelo de examen, y que merecen ser analizados y discutidos desde principios y criterios propios de una disciplina de evaluación.

(*) Universidad de Málaga.

Fecha de entrada: 31-01-2000 Fecha de aceptación: 12-04-2000

El examen tipo, que constituye el ámbito de referencia del presente estudio, se realiza por escrito, comporta un fuerte componente memorístico, puesto que al examinando no se le permite consultar libros o apuntes, y tiene un contenido estándar, aplicado a veces a centenares de alumnos o candidatos, en una o varias aulas a la vez. A ese prototipo obedecen, en España, no sólo la mayoría de los exámenes universitarios, sino también muchos exámenes en el Bachillerato, las pruebas de acceso a la universidad y algunos ejercicios en concursos-oposición de acceso a la función pública, en este caso, incluso, con un mismo formulario de prueba en distintas ciudades a la vez. Ese género de examen representa, pues, toda una institución y no solo en la universidad española.

Ser examinado es un modo especial de ser evaluado. Es una evaluación cuyos resultados repercuten, y a veces de modo decisivo, en la vida personal y no sólo en la carrera: en forma de ser aprobado o no, de obtener o no una plaza, un título, de alcanzar tal o cual puntuación que, a su vez, computará al lado de otras puntuaciones y méritos. El examen, y la correspondiente calificación, constituyen un filtro institucional de suma relevancia, en una sociedad que dice ser meritocrática y no discriminativa, excepto por las capacidades y los méritos.

La exacta delimitación del tema requiere puntualizar que existen muchos modos y ámbitos de evaluación, además del examen. Pueden y deben ser objeto de evaluación los procesos de enseñanza y no sólo los resultados de aprendizaje. Además es posible y necesaria la evaluación de los procesos de aprendizaje, de adquisición de destrezas, y no únicamente los productos de nuevos conocimientos. En fin, es recomendable realizar la evaluación de los alumnos sin objetivos de criba, ni de calificación, con fines de información para ellos mismos –alcanzable también mediante prácticas de autoevaluación– y para los docentes. Los exámenes, en suma, constituyen una forma particular de evaluación. Son, además, una forma pedagógicamente cuestionable –una cosa es aprobar, otra aprender–, no precisamente favorita de los estudiosos en sociología y teoría de la educación y en metodología evaluativa (Castells, 1989; Stufflebeam y Shinkfield, 1985, 1987; Wilson, 1992), los cuales suelen pasar de largo ante el examen tipo, aunque, por otra parte, reconozcan que nada influye tanto en los modos de estudio de los alumnos como el modo en que saben que van a ser examinados. Por cuestionable que sea, el caso es, pues, que el modo de examen:

- pasa a ser, retroactivamente, un factor influyente, crucial, en el proceso de enseñanza y aprendizaje, formando parte así de la actividad misma de instrucción;
- llega incluso a ser un factor que modifica (eventualmente, deforma) el núcleo del currículo o plan de estudios establecido; y, en consecuencia,
- es importante identificar las mejores formas prácticas, no ya sólo de evaluar, sino de examinar (Biggs, 1999; Gettinger, 1988; Lazarus, 1993).

Todo lo anterior vale, también, para la enseñanza universitaria. También en ella hay que examinar y no sólo evaluar, aunque ésa sea una de las tareas que los profesores juzgan menos gratificantes (Gros y Romaña, 1995).

Para el investigador familiarizado con la tecnología de las pruebas psicométricas, resulta tentador abordar el asunto en términos semejantes al de la fiabilidad y validez de los tests. Al fin y al cabo, se trata de una variedad de éstos: son tests de conocimiento. Cabe entonces incorporar también, por consiguiente, elementos de análisis psicométrico, como los de la teoría de

la respuesta al item (así, Rivas, Jornet y Suárez, 1995). Sin embargo, la específica naturaleza y las consecuencias de los exámenes académicos o selectivos, frente a los tests psicométricos, les constituyen en una especie muy singular de prueba, donde también es preciso idear y practicar procedimientos específicos de recogida de información pertinente, en orden a establecer su fiabilidad y validez.

Un enfoque de validez predictiva, o de criterio, respecto a un futuro profesional, salvo que éste se tome a largo plazo, podría incurrir en circularidad. Exámenes y expediente académico contribuyen a crear el futuro inmediato de los estudiantes y no meramente a predecirlo. Dado el peso que las calificaciones suelen tener en los inicios de la carrera laboral de cualquier graduado, habría que esperar algunos años para poder utilizar los logros profesionales como criterio de referencia para las calificaciones, y aun, entonces, lo sería para el promedio de éstas, y no para la nota académica en tal o cual materia.

La validez de un examen tiene que ver, ante todo, con la apropiada representación muestral del dominio de conocimientos que intenta someter a prueba. Sin embargo, no es fácil establecer con rigor la validez de contenido de un examen, a partir de su carácter de muestra representativa, extraída de algún universo o corpus de conocimientos que el alumno debe poseer. Salvo que ese corpus esté perfectamente acotado en un catecismo, una lista o un manual, carece de sentido atenerse a ese proceder, que si acaso vale al poner a prueba los conocimientos del alumno, en conjuntos claramente delimitados, como una tabla, un léxico o un código civil. Ahora bien, ninguno de los tramos de la enseñanza, y mucho menos la universitaria, se propone aleccionar sólo en tales géneros de conjuntos cerrados; como mínimo, pretende también instruir en cómo entenderlos y manejarlos. Qué debe conocer un examinando suele formar parte de la clase de las categorías difusas (Rosch, 1978). Un dominio, quizá, con núcleos prototípicos indudables, pero que se van difuminando hacia la periferia en contornos un tanto borrosos, lo que suele servir, por otra parte, para graduar la calificación entre el mero aprobado y el sobresaliente.

Para hacerla todavía más compleja, la confección de pruebas académicas fiables, válidas y –si posible fuera– normalizadas, encuentra una dificultad añadida. Un modelo de examen, con tales o cuales contenidos concretos, apropiadamente validado al modo usual de los tests psicométricos, sólo puede utilizarse una vez, en una convocatoria. Por su mero uso en una ocasión, queda invalidado para la siguiente, puesto que los examinandos saben o sospechan de antemano acerca del contenido del mismo. Así, pues, mientras que las pruebas psicométricas han de ser validadas y normalizadas también en su concreto contenido de unos elementos determinados, las pruebas académicas y las selectivas necesitan de validación, más bien en cuanto a formato y a tipo de ítem, pero no normalizadas en cuanto a contenido concreto, que suele quedar fuera de uso con una sola aplicación. En la validación de un examen –aparte de su pertinencia a un dominio acotable, aunque quizá difuso, de conocimiento– pasa a primer plano el formato de prueba. Tampoco este rasgo es exclusivo de los exámenes. En realidad, en toda clase de pruebas, y no sólo en las académicas o en aquéllas donde se comprueba el grado de instrucción, conocimiento y destrezas de los sujetos, el formato puede llegar a ser tan importante como el contenido. Cronbach (1998, p. 202) refiere que las puntuaciones, en formas de presentación similares, a veces correlacionan entre sí en grado más alto que las obtenidas en contenidos similares con formas diferentes. La adecuación de una prueba, por consiguiente, está ligada a cuestiones de formato, que son las consideradas a continuación.

Al prestar atención al formato de examen y al modo de su calificación, está justificado colocar el énfasis en una deseable cualidad, no ajena a la validez, y que en todo caso resulta crucial en los exámenes, cuando de las calificaciones depende el futuro de los examinandos. Es la cualidad de objetividad, imparcialidad y equidad, tanto de la prueba como de su calificación. Asumiendo un principio moral de mérito –«a cada uno según sus merecimientos»–, es lo que ante todo suele pedir y tiene derecho a exigir el examinando o aspirante a ser seleccionado, dentro de un sistema competitivo: que examen y calificación no sean arbitrarios, no estén sujetos a caprichos, humores o preferencias personales del examinador. La objetiva equidad de un examen y de su calificación es seguramente el más relevante elemento constitutivo de su fiabilidad y validez.

Para poder cumplir con el requisito de equidad, cuando es muy elevado el número de sujetos que han de pasar una prueba, y quedan fuera de consideración otras fórmulas posibles (como el acuerdo entre varios jueces), en muchas pruebas académicas o de selección es común acudir a la fórmula llamada «prueba objetiva» o examen «tipo test» que luego, además, puede ser corregido mediante lectora óptica. Consiste en una lista, más o menos extensa, de ítems, en bloques de tres o más enunciados (hay discusión sobre el número más apropiado: cf. Delgado y Prieto, 1998), donde el examinando ha de señalar cuál es verdadero entre otros falsos. Las principales reglas de construcción de estas pruebas de alternativas múltiples son conocidas desde hace tiempo (Berk, 1984, p. 227; Haladyna, 1994; Haladyna y Downing, 1989).

Si los ítems están bien formulados (lo que no es fácil, pero tampoco imposible), en pruebas de ese género parecen quedar a salvo la objetividad e imparcialidad, así como también la economía de tiempo de corrección, cuando son muchos los examinados. Presentan ventajas apreciables: comparar alternativas e identificar la respuesta correcta, genera menos ansiedad, en los examinandos que tener que construir una respuesta (Embretson, 1985). Hay estudios, además, que ponen de manifiesto la alta correlación que exhiben sobre todo con respuestas de resumen (correlación hasta 85), pero también, aunque no tan elevada, con una composición libre (Breland, 1979; Hogan y Mishler, 1980).

Sin embargo, ni siquiera con la ponderación (penalizadora) de las respuestas incorrectas, el formato de pruebas objetivas es capaz de eliminar las habilidades puramente adivinatorias –e irrelevantes para el conocimiento de la materia– que a algunos examinandos les permiten excelentes puntuaciones (cf. Cronbach, 1998, pp. 94-95). En todo caso, la principal amenaza a la validez de esas pruebas no está en la interferencia de estrategias adivinatorias, que son susceptibles de control (Budescu y Bar-Hillel, 1993; Prieto y Delgado, 1999). Está en que adolecen de dos graves sesgos que convertirían en perverso un sistema evaluador y selectivo basado sólo en ellas. El primero es que para responder correctamente no hace falta saber disertar sobre los temas y ni siquiera escribir. Son pruebas ágrafas, iletradas. A la postre y a la larga, con tales pruebas podrían llegar a maestros, inspectores, psicólogos, abogados o jefes administrativos, personas que no supieran redactar un informe ni poner por escrito un plan de trabajo. El segundo es que valen para proposiciones indiscutibles y en materias axiomáticas, pero no, o no tanto, para aquéllas donde el pensamiento crítico o el razonamiento forman parte esencial de la capacidad, destrezas y conocimientos adquiridos. Dicho en términos de psicología cognitiva: el examen tipo test mide ante todo pensamiento convergente, memoria de reconocimiento, destrezas muy específicas –o acaso trucos– para acertar en ese

tipo de prueba, y no otros conocimientos y capacidades.

El doble estudio, del que se informa a continuación, tiene que ver, no con cuestiones de contenido, de adecuado muestreo del dominio de conocimientos por poner a prueba, sino con otro elemento no menos relevante, como ya se ha razonado: el formato de la prueba. Está al servicio del diseño de modos de examen objetivos, imparciales, equitativos, ecuánimes, mas no ágrafos. Trata de ver si otros formatos, distintos del convencional de «prueba objetiva», pueden satisfacer igualmente esos requisitos sin por ello incurrir en los sesgos recién mencionados; si hay alguna alternativa a las pruebas objetivas, que sin perder objetividad y ecuanimidad, se halle libre de los riesgos de efectuar selección profesional, o de conceder titulación universitaria, a espaldas de una competencia discursiva que el examen tipo test es incapaz de captar y, en consecuencia, de fomentar retroactivamente.

El primer autor utiliza, desde hace años, un formato de examen que consiste en una lista de enunciados a semejanza de las pruebas objetivas, pero sin bloques, en proposiciones aisladas. En ellas, el examinando ha de empezar por identificar cuáles son verdaderas y cuáles falsas, pero además –y aquí reside lo peculiar del procedimiento– en las proposiciones falsas ha de sustituir los términos inapropiados o juicios erróneos, por otros que sean correctos. A menudo, esa sustitución puede hacerse de varios modos, todos ellos acertados, y también caben grados de calidad en el acierto, grados que serán objeto de calificación, ítem a ítem.

Este formato –se presume– mantiene ventajas propias de las pruebas objetivas: la de una alta objetividad en la calificación y también rapidez en la corrección. Y quizá no se expone a los inconvenientes de aquéllas: el hecho de que el examinando ha de formular por su cuenta juicios alternativos, cuando los ítems propuestos son falsos, contribuye a evitar los sesgos antes señalados y puede que oriente hacia modos más críticos y reflexivos de estudio. Los ítems son corregidos y puntuados uno a uno, en escala de cero a diez, y existe penalización cuando no hubo acierto en la identificación correcto/falso. En el Anexo se recogen las instrucciones típicas para este formato de examen y algunos ejemplos de ítems, así como de sus posibles formulaciones de respuestas certeras.

La finalidad concreta del estudio ha sido, pues, ver cómo funciona este formato de examen en comparación con el de las pruebas objetivas convencionales y con alguna otra modalidad de prueba, también tradicional, como es un conjunto de preguntas que requieren respuesta breve o el desarrollo algo más extenso de un tema. La comparación se hace sobre un constructo de «imparcialidad», «ecuanimidad» y «objetividad», entendida ésta como «no subjetividad», no intrusión en la valoración por parte de la persona que corrige el examen. A falta de otro criterio, ese constructo va a ser operacionalizado del siguiente modo: se postula como prueba más objetiva, imparcial, ecuánime (términos aquí equivalentes) aquélla que correlacione con las demás con valores más altos.

MÉTODO

Sujetos y procedimiento

En un primer grupo, 55 alumnos, con la titulación de Psicología, realizaron el examen de la materia, con tres formatos diferentes de prueba:

Prueba 1 (prueba objetiva), de tipo test, con bloques (20 en total) de tres enunciados, para reconocer y marcar uno de ellos como verdadero entre otros falsos *(test)*;

Prueba 2 (prueba con lista de enunciados singulares, 10 en total), a semejanza de la prueba objetiva, pero sin bloques,

donde era preciso identificar proposiciones falsas y sustituir en ellas términos inapropiados o juicios erróneos por otros correctos *(corrección)*;

Prueba 3 (preguntas breves, 5 en total) en las que se pedía una respuesta concisa y breve, máximo media docena de líneas *(preguntas)*.

En un segundo grupo, fueron 82 alumnos, con la titulación de Psicopedagogía, los que realizaron el examen con los anteriores procedimientos, más una cuarta modalidad: propuesta de dos temas, y elección de uno de ellos para ser desarrollado por el alumno, a manera de ensayo, con una extensión aproximada de un folio *(tema)*.

Los alumnos del primer grupo, dispusieron de un máximo de hora y media para realizar los tres ejercicios correspondientes a los distintos formatos de prueba. Los del segundo grupo, con un ejercicio más, dispusieron de dos horas.

Conviene informar acerca de algunas otras circunstancias relevantes, aunque no esenciales, para los fines de la investigación. Los examinandos podían consultar libros y apuntes. Los contenidos concretos de los diversos formatos de examen tenían que ver más bien con la comprensión crítica de los textos estudiados, con la inferencia y el razonamiento, más que con la memorización. Por otra parte, los alumnos habían sido informados, desde principio de curso, sobre la estructura múltiple del examen final. Dos semanas antes de éste, realizaron un ejercicio de evaluación «a modo de examen» para familiarizarse sobre todo con el formato de la prueba con lista de enunciados singulares, y evitar efectos de sorpresa o desorientación, pero también con los formatos 1 y 3, es decir el de la prueba objetiva y el de las preguntas breves, y contribuir así a paliar posibles diferencias entre ellos que pudieran deberse a la experiencia anterior con tales modalidades de prueba.

Evaluación de los ejercicios

La prueba 1 fue puntuada con lectora óptica. Las pruebas 2, 3 y 4 fueron corregidas por el profesor, sobre la base de criterios de valoración previamente establecidos y «a ciegas» en el siguiente sentido: sin ver el nombre de cada examinando, y juzgando cada prueba, por separado, en días distintos, e ignorando, al calificar una prueba, la calificación obtenida en otra u otras pruebas ya corregidas anteriormente.

La calificación que recibieron los alumnos fue la media de las tres o cuatro puntuaciones, según la pertenencia al primer o al segundo grupo, respectivamente. Interesa señalar que no se formuló objeción o queja alguna acerca del planteamiento o de los contenidos del examen, ni antes ni después del mismo. En una sesión posterior, los alumnos tuvieron acceso tanto a su propio ejercicio, cuanto a un ejercicio «modélico» (ver Anexo), por así decir, confeccionado por el profesor como referencia de contraste. Tras esa sesión ningún alumno solicitó revisión de su calificación.

RESULTADOS

Las imágenes visuales de la distribución de las puntuaciones, de los sujetos, se obtienen mediante los diagramas de dispersión de las mismas. En cada diagrama sólo pueden representarse dos escalas de puntuación a la vez. Como muestra se ha elegido presentar, en gráfico I, el espacio definido por las puntuaciones del grupo 2, en las dos modalidades en principio más afines: la de «test» y la de «corrección». En el diagrama cabe observar que se dibuja el perfil de una cierta asociación entre los resultados en ambas modalidades. Pero, por otra parte, aparecen no pocos puntos desperdigados, fuera de perfil, algunos incluso de modo extremo, en solitario. Así, pues, para unos pocos sujetos el resultado

GRÁFICO I

Dispersión de puntuaciones

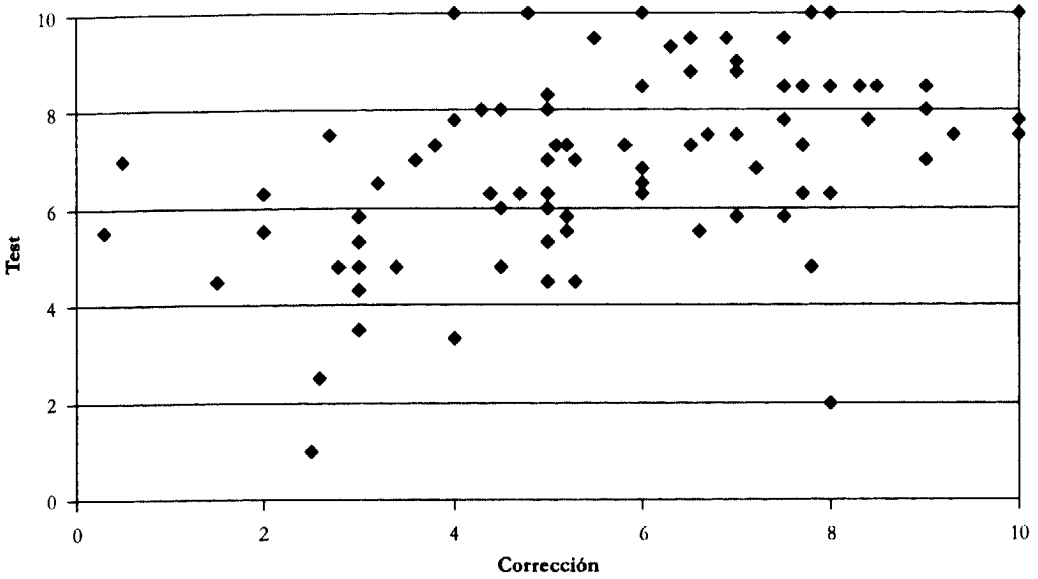

Diagrama de dispersión de puntuaciones en las modalidades de examen y de corrección (grupo 1; n = 82)

ha sido completamente dispar de una modalidad a otra, a veces como para diferir en calificación de suspenso a notable. Diagramas semejantes surgen al ubicar los datos de otras parejas de puntuaciones. En ellos aparecen perfiles análogos, pero también puntos aislados, correspondientes a discrepancias grandes para un mismo sujeto, entre sus logros en una modalidad y en otra. Éste es el resultado de mayor relieve a través de los distintos diagramas de dispersión obtenidos, que, sin embargo, no merece la pena reproducir uno por uno. Más allá de una inspección intuitiva de gráficos, es el análisis correlacional el que permite realmente hacerse cargo de las asociaciones entre variables.

Los resultados del análisis correlacional proporcionan una imagen ya no visual, sino abstracta, pero clara, de las asociaciones entre las puntuaciones en las distintas pruebas. Se han obtenido las correlaciones simples y parciales entre esas puntuaciones, y, asimismo, entre cada una de éstas con la media de las restantes y con la media total.

La tabla I presenta la matriz de esas correlaciones. En la zona superior de cada celda aparecen los valores de correlación en el estudio con el primer grupo de alumnos; en la inferior, los del segundo grupo. Por otro lado, a la izquierda de la barra inclinada (/) están los valores de la correlación de Pearson; a la derecha, los de la respectiva correlación parcial entre las variables respectivas, tras eliminar la asociación explicable por la otra variable (en grupo 1) o las otras dos variables (en grupo 2) de los demás formatos de examen.

TABLA I

Matriz de correlaciones

	Test	Corrección	Preguntas	Tema
Corrección	.38* /.19 .58* /.18			
Preguntas	.57* /.48* .58* /.45*	.43* /.28 .52* /.35*		
Tema	– .30* /.19	– .32* /.22	– .18 / –.07	
Media restantes	.46* .59*	.56* .57*	.60* .55*	– .32*
Media	.82* .79*	.76* .79*	.82* .77*	– .57*

Los resultados son consistentes de un grupo a otro. La modalidad «tema» queda en los valores más bajos de relación con las demás modalidades. En las correlaciones simples, los otros tres formatos de prueba intercorrelacionan con valores muy semejantes entre sí y, como no podía ser menos, con la nota media. Estos valores son ligeramente superiores en las modalidades «test» y «preguntas» frente a «corrección», en el primer grupo; mientras que en el segundo, los valores de correlación simple de estas tres variables son prácticamente idénticos.

El perfil de las correlaciones parciales tiene alguna particularidad concretada en el formato «pregunta». Esta variable conserva valores altos y significativos en su asociación con los formatos «corrección» y «test», mientras que desciende a casi valor cero, pero negativo, en su correlación con «tema».

Se han efectuado también, para ambos grupos, los correspondientes análisis de regresión, tomando la puntuación media como variable dependiente y las puntuaciones de modalidad como independiente. Los resultados de este análisis proporcionan otra perspectiva aunque, como es obvio, sobre un mismo paisaje. Los pesos beta ponderados, de la ecuación de regresión, han sido de .43 (primer grupo) y .35 (segundo grupo) para la variable «test»; de .42 y .37, respectivamente, para «corrección»; y de .40 y .33 para «preguntas». Los pesos beta descienden, del primer al segundo grupo, por la introducción en éste de la puntuación en «tema», cuyo peso beta es de .30.

Se procedió, en fin, al análisis factorial de los resultados, al análisis de los componentes principales. En ambos grupos pudo extraerse un único factor, por lo que no hubo lugar a buscar soluciones rotadas. El porcentaje de varianza explicado por ese factor único, fue de un 64,2%, en el grupo 1 y de un 55,3%, en el grupo 2. La tabla II expone la matriz de componentes, por modalidades de examen, para cada uno de los grupos.

TABLA II

Análisis factorial

	Grupo 1	Grupo 2
Corrección	.732	.791
Preguntas	.844	.804
Tema		.529
Test	.893	.812

Resultados del análisis de componentes principales: solución de un solo factor sin posibilidad de rotación

Lo mismo que sucede en otros análisis, en éste, de componentes principales, la modalidad «tema» queda netamente por debajo de las otras, que en cambio aparecen con valores semejantes. Los valores más altos, con todo, corresponden al formato «test» y los más bajos al de «corrección».

DISCUSIÓN

A través de todos los análisis, tres de los tipos de examen se presentan con valores parecidos: «test», «preguntas», «corrección». En principio, pueden considerarse aproximadamente igual de ecuánimes. Si el desafío era mostrar que el formato de «corrección» es tan equitativo y objetivo como el de «test», su equivalencia a ese efecto ha quedado probada. Un resultado no previsto es que el formato de «preguntas» breves, aun siendo pocas (cinco), no se queda por debajo en esa virtud, antes bien, sobresale. Lo que menos cabía esperar es que, en las correlaciones parciales, ese formato presenta valores significativos y más altos con «corrección» y aun más elevados con «test», alcanzando aquí valores de .45 y .48, mientras que la correlación de estos otros formatos, más cercanos entre sí por contenido, cae a niveles que la tornan no significativa.

Es significativo el dato de que el formato «preguntas» tenga correlación más alta con «test» que con «corrección». Pero el resultado más intrigante está en la correlación tan baja de «preguntas» con «tema», cuando ambos formatos, más que los otros, poseen un elemento en común, que además es del todo ajeno al «test»: la exigencia de que el examinando muestre que es capaz de redactar unos fragmentos de discurso científico. Tan inesperado resultado no es interpretable desde las bases y los hallazgos del presente estudio. Su confirmación, primero, y su interpretación, después, han de aguardar a otros estudios que acoten mejor y se centren de manera específica en el formato «preguntas», para poder examinar de modo sistemático sus asociaciones con los formatos de contenido más contrapuesto: el de tipo «test» y el de «tema».

Entretanto, sí que pueden extraerse algunas conclusiones sólidas respecto al objetivo principal del estudio: respecto a la imparcialidad, ecuanimidad, objetividad –en el sentido aquí manejado– de los distintos modos de prueba:

- Los exámenes tipo test, los de preguntas breves y los de corrección de frases incorrectas son,

aproximadamente, igual de ecuánimes, objetivos e imparciales; mejor dicho, pueden serlo: lo han sido en el presente estudio, y es razonable esperar que oportunos refinamientos que se introduzcan en ellos, contribuyan a mejorarlos, pero, previsiblemente, en mejora paralela, sin que alguno llegue a despegar mucho respecto a los demás.

- Una calificación final extraída a partir de distintos tipos de prueba siempre será más ecuánime que aquélla que se derive de un solo tipo. Sin embargo, no cabe desechar, como injusto o sesgado, el uso de una modalidad única de prueba entre las tres aquí igualadas: «test», «corrección», «preguntas». Las correlaciones que cada una de ellas obtiene, en los dos grupos, con la que aquí puede servir de variable criterio –la puntuación final o combinada, de promedio– alcanzan valores lo bastante elevados (entre .76 y .82) como para poder afirmar que cumplen bien con su función, ordenar en su nivel de rendimiento a un conjunto de examinandos, de calificarlos sin injusticia y, en consecuencia, de cumplir, en su caso, una función social de selección de candidatos.
- Sin embargo y por desgracia, la ecuanimidad conseguida, respecto al conjunto de sujetos, no puede generalizarse a todos y cada uno de éstos, tomados uno a uno. Como aparece en los diagramas de dispersión de las puntuaciones (gráfico I), algunos sujetos concretos, de ser evaluados y calificados por uno u otro formato, llegan a oscilar nada menos que del notable al suspenso. Permanece, pues, pendiente la cuestión de la ecuanimidad de los exámenes y pruebas de selección con respecto a los individuos.

Quedan abiertas numerosas cuestiones que no es posible resolver, a partir del diseño y de los datos de este estudio: qué perfil de resultados se hubiera obtenido de no haber conocido los sujetos, de antemano, el modo o modos en que iban a ser evaluados; qué mejora en las correlaciones del formato «tema» podría haberse logrado con una corrección por varios jueces; hasta dónde se pueden generalizar los resultados a otras situaciones de prueba, tales como la valoración de conocimientos, en el primer ejercicio, para la obtención del carné de conducir, o las de acceso de candidatos a la función pública, realizadas a veces con grupos multitudinarios; qué sentido tiene seleccionar, principalmente, sobre la base de un saber más bien teórico y libresco, cuando, cada vez más, la selección de personal presta atención, por una parte, a destrezas prácticas y, por otra, a variables de personalidad (Borman, Hanson y Hedge, 1887; Hogan y Roberts, 1996). Mucho menos cabe dar respuesta a otras cuestiones suscitadas por la influencia retroactiva del tipo de examen sobre el aprendizaje: cuáles son las consecuencias positivas y negativas de la aplicación de tal o cual tipo de prueba; y, sobre todo, cuáles son, a largo plazo, y por efecto acumulativo, sobre sucesivas cohortes de estudiantes. Pero como acostumbra a decirse, son ya otras cuestiones, es ya otra historia, a la que el presente estudio no podía atender.

BIBLIOGRAFÍA

BERK, R. A. (ed.): *A guide to criterion-referenced test construction.* Baltimore, John Hopkins Univ. Press, 1984.

BIGGS, J.: *Teaching for quality learning at University.* Ballmoor, Open University Press, 1999.

BORMAN, W. C., HANSON, M. A. y HEDGE, J. W.: «Personnel selection», en *American Review of Psychology,* 48 (1997), pp. 299-347.

BRELAND, H. M.: *Can multiple-choice tests measure writing skills?* Nueva York, College Entrance Examination Board, 1979.

BUDESCU, D. y BAR-HILLEL, M.: «To guess or not to guess», en *Journal of educational measurement,* 14 (1993), pp. 197-201.

CASTELLS, M.: «Los sistemas de evaluación de las Universidades», en VARIOS AUTORES: *Hacia una clasificación de las Universidades según criterios de calidad.* Madrid, Consejo Universidades / Fundación Universidad-Empresa, 1989.

COLLIS, K. y ROMBERG, T. A.: «Evaluación del desempeño en matemáticas: un análisis de item de pruebas abiertas», en M. C. WITTROCK y E. L. BAKER (eds.): *Test y cognición,* Barcelona, Paidós, 1998.

CRONBACH, L. J.: *Essentials of psychological testing / Fundamentos de los tests psicológicos.* Londres/Madrid, Harper Collins / Biblioteca Nueva, 1998.

DELGADO, A. R. y PRIETO, G.: «Further evidence favoring three option in multiple-choice tests», en *European Journal of Psychological Assessment,* 3, 14 (1998), pp. 197-201.

EMBRETSON, S. (ed.): *Test design: Development in Psychology and psychometrics.* Orlando, Fl., Academic Press, 1985.

FERNÁNDEZ-VALLINA, J.: «Selección MIR, BIR, FIR, QIR», en VARIOS AUTORES: *Hacia una clasificación de las Universidades según criterios de calidad.* Madrid, Consejo Universidades/Fundación Universidad-Empresa, 1989.

GETTINGER, M.: «Analogue asessment: Evaluating academicabilities», en E. S. SHAPIRO y T. R. KRATOCHWILL (eds.): *Behavioral assessment in schools.* Nueva York, Guilford, 1998.

GROS, B. y ROMAÑA, T.: *Ser profesor: palabras sobre la docencia universitaria.* Barcelona, Univ. Barcelona, 1995.

HALADYNA, T. M.: *Developing and validating multiple-choice test items.* Hillsdale, N.J., L. Erlbaum, 1994.

HALADYNA, T. M. y DOWNING, S. M.: «A taxonomy of multiple-choice item-writing rules», en *Applied measurement in Education,* 2 (1989), pp. 37-50.

HOGAN, T. P. y MISHLER, C.: «Relationships between essay tests and objective tests of language skills for elementary school students», en *Journal of Educational Measurement,* 17 (1980), pp. 219-227.

HOGAN, R. H.; HOGAN, J. y ROBERTS, B. W.: «Personality measurement and employment decisions», en *American Psychologist,* 51 (1996), pp. 469-477.

LAMO DE ESPINOSA, E.: «Evaluación de la calidad de la enseñanza», en VARIOS AUTORES: *Hacia una clasificación de las Universidades según criterios de calidad.* Madrid, Consejo Universidades / Fundación Universidad-Empresa, 1989.

LAZARUS, B.: «Best practices in assessing academic achievement», en H. B. VANCE (ed.): *Best practices in assessment for school and clinical settings.* Brandon, Clinical Psychology Publishing, 1993.

LENTZ, F. E.: «Direct observation and measurement of academic skills: a conceptual review», en E. S. SHAPIRO y T. R. KRATOCHWILL (eds.): *Behavioral assessment in schools.* Nueva York, Guilford, 1998.

O.C.D.E.: *Escuelas y calidad de la enseñanza.* Madrid/Barcelona, M.E.C./Paidós, 1991.

PRIETO, G. y DELGADO, A. R.: «The effect of instructions on multiple-choice test scores», en *European Journal of Psychological Assessment,* 2, 15 (1999), pp. 143-150.

RIVAS, F.; JORNET, J. y SUÁREZ, J. M.: «Evaluación del aprendizaje escolar», en F. SILVA (ed.): *Evaluación psicológica en niños y adolescentes.* Madrid, Síntesis, 1995.

ROSCH, E.: «Principles of categorizations», en E. ROSCH y B. B. LLOYD (eds.): *Cognition and categorization.* Hillsdale, N. J., L. Erlbaum, 1978.

SCHÖN, D. A.: *La formación de profesionales reflexivos.* Madrid/Barcelona, MEC / Paidós, 1992.

STUFFLEBEAM, D. L. y SHINKFIELD, A. J.: *Evaluación sistemática: guía teórica y práctica.* Madrid/Barcelona, MEC / Paidós, 1987.

TYLER, R. W.: «General statement on evaluation», en *Journal of educational research,* 35 (1942), pp. 492-501.

WILSON, J. D.: *Cómo valorar la calidad de la enseñanza.* Barcelona/Madrid, Paidós / MEC 1992.

ANEXO

EJEMPLOS DE ÍTEMS EN EL FORMATO «CORRECCIÓN»

1. La Psicología se ha interesado más por aspectos funcionales que por aspectos estructurales del comportamiento.
2. Aprendizaje es todo cambio de conducta en un individuo.
3. Hay en América más personas que hablan el castellano que en Europa.

EJEMPLOS DE SU POSIBLE «CORRECCIÓN»

1. La Psicología *no* se ha interesado *apenas* [más] por aspectos estructurales [que] *y sí, casi exclusivamente,* por aspectos funcionales del comportamiento.
2. Aprendizaje es todo cambio *en el potencial* de conducta en un individuo *como consecuencia de la práctica o de la experiencia.*
3. *(es verdadera)*