JOSÉ VICENTE ROMEU

«PRUEBAS OBJETIVAS» EN LOS EXAMENES DE GEOGRAFIA E HISTORIA. RESULTADOS *

INTRODUCCIÓN

Como continuación a nuestro trabajo en el número XIV de esta misma Revista [1], en el que informábamos acerca de la investigación iniciada en el Seminario de Psicología de la Facultad de Filosofía y Letras de Valencia, al objeto de preparar y validar cuestionarios de "exámenes objetivos" en las materias de Geografía e Historia, se nos brinda ahora la oportunidad de presentar los resultados y las conclusiones principales de nuestra tesis.

Por si esclarece mejor la motivación de nuestro trabajo, quisiéramos insistir en que esta investigación se llevó a cabo no con el único interés de satisfacer una curiosidad intelectual, sino más bien con la idea de prestar un servicio concreto a la docencia universitaria, cual es el del perfeccionamiento de las técnicas del control del aprovechamiento escolar.

En nuestra tesis, tras la presentación y planteamiento de una serie de problemas generales a propósito de la medida en Psicología y de un modo más específico de la medida del aprovechamiento escolar, así como tras la presentación de los llamados tests de aprovechamiento o "Pruebas objetivas", con una breve consideración histórica y algunas consideraciones acerca de sus ventajas y desventajas, pudimos fijar nuestros propósitos en los siguientes términos:

1) Ver hasta qué punto es posible sustituir el examen tradicional por las llamadas "pruebas objetivas", concretamente en Geografía e Historia en los cursos comunes de esta Facultad.

* Extracto de la tesis doctoral presentada por su autor en junio de 1966, en la Facultad de Filosofía y Letras de la Universidad de Valencia.

[1] ROMEU, J. Vicente: *Pruebas objetivas en los exámenes de Geografía e Historia.* "Saitabi", XIV (1964), pp. 239-43.

2) Considerar las ventajas de todo tipo que tal sistema de examen ofrece tanto al alumno como al profesor, sobre todo si con ellos se podía emitir un juicio más exacto en cuanto a la capacitación y aprovechamiento del alumno.

3) Aunque nuestro propósito fuera tan concreto, es indudable que nuestros resultados podrían servir también de orientación en cuanto a posibles aplicaciones a otras asignaturas y cursos.

EXAMEN-ENSAYO

Antes de iniciar nuestro trabajo, a modo de pretesting, se realizó un ensayo en los exámenes de septiembre de 1964, en la asignatura de Geografía. Nuestra experiencia consistió en añadir a la forma de examen tradicional otro, en forma de "prueba objetiva". Las condiciones en que nos tuvimos que mover no fueron las ideales, esto es, la reducida muestra de 49 alumnos era además sesgada dado el carácter de examen extraordinario, parte de ellos no se habían presentado y parte habían sido suspendidos en junio. La "prueba objetiva" con un total de 40 ítems (18 de completamiento y 22 de múltiple opción) se preparó sobre dos de los temas señalados para el examen. Por otro lado, los alumnos fueron invitados a responder por escrito a otros dos temas.

Corregidas ambas pruebas por personal técnico y con las debidas precauciones, se procedió a la tipificación de las puntuaciones, al objeto de poder efectuar las comparaciones pertinentes.

Al estudiar la fiabilidad de las pruebas obtuvimos los resultados siguientes:

Examen tradicional	$r = .14$	S. B. $= .23$
Prueba objetiva	$r = .52$	S. B. $= .68$

Esto nos indicaba ya una mayor justeza en los juicios que se podían emitir partiendo de la "prueba objetiva".

En cuanto a la validez, al correlacionar las puntuaciones típicas obtenidas por los alumnos en ambas formas de examen, obtuvimos un coeficiente de Pearson de $r = .43$, índice no muy valioso, aunque dada la baja fiabilidad del criterio no cabía esperar otra cosa. Por otro lado, téngase en cuenta también que realmente no suelen obtenerse en pruebas similares índices superiores a $r = .65$ [2].

A pesar de ello, en la siguiente tabla de expectancia puede ponderarse la capacidad predictiva de esta "prueba objetiva" respecto del examen tradicional.

[2] J. STANLEY AHMAN: *Testing Student Achievement and Aptitudes*, p. 54.

Tabla de expectancia para las puntuaciones del examen tradicional y de la "prueba objetiva"

Puntuaciones en el examen tradicional	*Puntuaciones en la "prueba objetiva"*			
	Inferiores	*Promedio*	*Superiores*	*Total*
Superiores	0	4	3	7
Promedio	4	18	4	26
Inferiores	6	8	2	16
Total	10	30	9	49

Donde puede verse que de los 39 sujetos que alcanzan o superan el promedio de la "prueba objetiva", 29 también lo alcanzan o superan en el examen tradicional, lo cual equivale a un 66% de capacidad predictiva; mientras que sólo 4 de los sujetos considerados inferiores al promedio en las calificaciones de la "prueba objetiva" alcanzaron puntuaciones que estaban en el promedio del examen tradicional, sin que ninguno de ellos superase el promedio.

Ahora bien, dadas las limitaciones e irregularidades de las condiciones en que habíamos ensayado, así como la baja fiabilidad de las pruebas usadas, consideramos que el índice de validez obtenido era de gran significación, por todo lo cual supusimos que, si las "pruebas objetivas" se preparaban sobre un muestreo más amplio del contenido de la materia de examen y se usaban técnicas precisas tanto para su preparación como para su corrección y puntuación, indudablemente aumentaría la fiabilidad de las mismas y su validez, en parte también, como consecuencia del aumento de su fiabilidad.

Plan de trabajo

Nuestro trabajo se planificó del modo siguiente:

a) Construcción de las pruebas objetivas que habían de utilizarse en los exámenes de Geografía e Historia, de modo que su contenido tuviera una cierta representatividad muestral en el conjunto de conocimientos que debían suponerse en un determinado nivel de instrucción.

b) Establecimiento de las normas a seguir en cuanto a la forma en que las pruebas debían ser redactadas, al formato de su presentación y a las técnicas de aplicación, corrección y puntuación de las mismas.

c) Elaboración estadística de las puntuaciones y estudio de la fiabilidad y validez de las pruebas usadas, así como análisis interno de las mismas, al objeto de ulteriores aplicaciones.

PRUEBAS USADAS

Las pruebas elaboradas y usadas por nosotros se administraron al total de la muestra de alumnos de cada materia a lo largo del curso, como se indica en el cuadro siguiente:

Prueba	*Fecha*	*Curso*	*Núm. de alumnos*
Geografía 1	Enero 1965	2.º Comunes	173
Geografía 2	Abril 1965	2.º Comunes	167
Geografía 1 bis	Diciembre 1965	2.º Comunes	239
Historia A	Enero 1965	1.º Comunes	247
Historia B	Enero 1965	1.º Comunes	239

Tales pruebas constaban de diversos tipos de ítems y su extensión era relativamente proporcional a la materia acotada para cada examen. En el cuadro siguiente se presentan estos pormenores.

Cuadro general de las "pruebas" objetivas" con expresión de las fechas en que se aplicaron, del tiempo concedido y de la composición de las mismas.

Prueba	*Fechas*	*Tiempo*	*Compl.*	*Asoc.*	*Múlt. opc.*	*C. Brev.*	*Total*
Geografía 1	Enero 1965	1 h. 30'	22	20	19	14	75
Geografía 2	Abril 1965	45'	20	20	20	—	60
Geografía 1 bis	Dicbre. 1965	1 h.	20	20	20	4	60
Historia A	Enero 1965	1 h.	12	25	30	7	74
Historia B	Enero 1965	1 h.	12	22	30	7	71

Para el muestreo de los ítems así como para la forma en que fueron redactados se contó siempre con el asesoramiento de los señores catedráticos encargados de las asignaturas respectivas, tratando de lograr que el conjunto de los ítems contenidos en cada prueba fuera lo más representativo posible del total de la materia acotada para el examen. Esto era de gran importancia, puesto que tal representatividad había de constituir para nosotros uno de los criterios de validez de mayor estima.

APLICACIÓN, CORRECCIÓN Y PUNTUACIÓN DE LAS PRUEBAS

Al objeto de lograr la mayor eficiencia posible, se redactaron normas relativas a la forma en que debía disponerse al alumnado y cómo debía

efectuarse el pase de las pruebas. Se dieron instrucciones precisas y se fijó el tiempo de realización.

Para facilitar la corrección se prepararon claves con las respuestas exactas y se redactaron respuestas esquemáticas que sirvieran como criterio para juzgar las llamadas cuestiones breves.

La puntuación directa se obtuvo por la suma de aciertos (PD. =A). Puesto que en las llamadas cuestiones de completamiento pueden darse dos o más claros, se hizo un estudio comparativo entre las posibles formas de puntuación (una puntuando las cuestiones por partes y otra en conjunto) y se halló que las diferencias no eran significativas. La correlación de Pearson entre ambas formas de puntuación dio un índice de r = .93. Al objeto de corregir los aciertos por azar en las llamadas cuestiones de múltiple opción, se aplicó la conocida fórmula

$$PD. = A - \frac{E}{n-1}$$

a la que se añadía una constante para evitar posibles puntuaciones negativas. Para puntuar las cuestiones breves, se usó una escala de 0 a 3, al objeto de que pudiera matizarse mejor el juicio valorativo.

Para cada prueba se preparó una tabla con las normas de puntuación, tanto para sus diversas partes como para el total, con expresión de las puntuaciones máximas y mínimas alcanzables.

Los datos fueron recogidos en fichas individuales, al objeto de facilitar el estudio y elaboración estadística de los mismos.

Tratamiento estadístico de las puntuaciones

Recogidas las puntuaciones alcanzadas por los sujetos en cada una de las pruebas, procedimos al tratamiento estadístico de las mismas, al objeto de convertirlas en otras de carácter estandardizado, con las que poder operar con mayor rigor y hacer las comparaciones pertinentes.

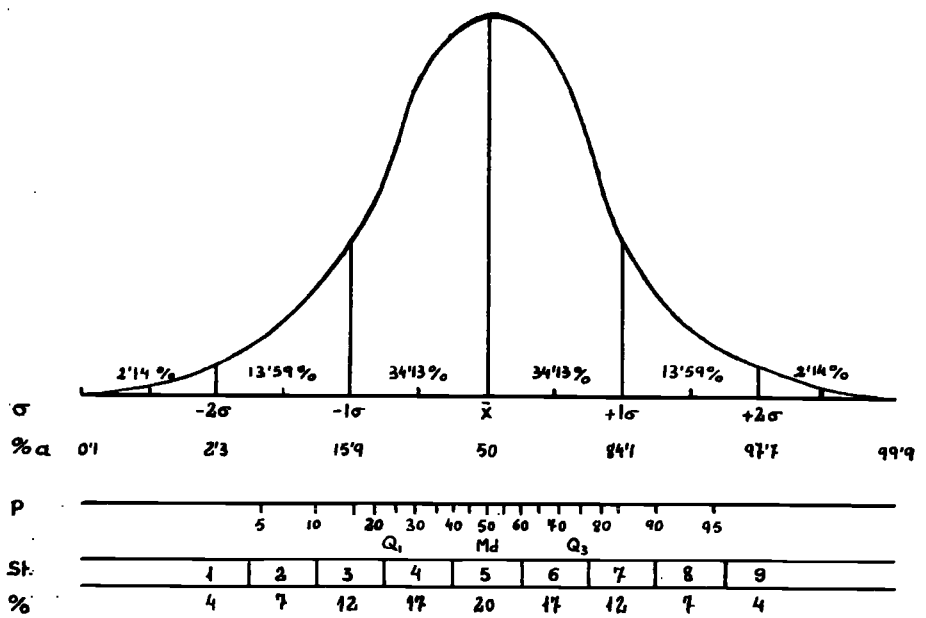
Las puntuaciones directas (PD.) apenas sirven para discriminar entre los sujetos. Para interpretar dichas puntuaciones es preciso compararlas con las obtenidas por los demás sujetos de la muestra. Es decir, que las puntuaciones directas han de ser interpretadas según se aproximen o separen más de la medida del grupo normativo. Quiere decirse que hemos de averiguar la posición exacta que cada P.D. ocupa en el grupo normativo, y en esto es en lo que consiste precisamente la tipificación, en transformar las P.D. en otras puntuaciones (típicas o standard) que de algún modo nos indiquen esta posición.

Para ello procedimos a elaborar tablas de distribución de frecuencias de las puntuaciones obtenidas por los sujetos en cada una de las pruebas,

tanto en el total como en cada una de sus partes. De cada una de ellas trazamos su representación gráfica al objeto de observar su normalidad. Obtuvimos luego en cada caso los principales estadísticos: media, mediana y sigma (desviación típica o standard). Finalmente, en un estudio crítico de las mismas, calculamos los errores típico [3] y muestral [4] de cada una de ellas.

Aunque en nuestro caso, las muestras usadas no se pudieran prejuzgar como "población", tanto su amplitud como la distribución aproximadamente normal de sus puntuaciones, nos permitieron trabajar con ellas como si fueran grupos normativos, bien que teniendo en cuenta los errores muestrales obtenidos a un nivel de confianza del 5 por 100.

Frecuentemente las puntuaciones típicas son expresadas en términos (centiles) que nos indican el tanto por ciento de sujetos de la población normativa a que es superior un determinado sujeto de ella. Sin embargo, estas puntuaciones presentan un grave inconveniente, y es que su unidad no es constante. A nosotros nos pareció más conveniente usar las llamadas



Relación entre puntuaciones típicas, percentiles y estaninos.

---

[3] $\sigma_{\bar{X}} = \dfrac{\sigma}{\sqrt{N-1}}$; $\sigma_\sigma = 0'707\ \sigma_{\bar{X}}$; $\sigma\,\text{Med.} = 1'25\ \sigma_{\bar{X}}$

[4] 1'96 . error típico.

puntuaciones estaninas (o eneatipos), las cuales a su vez nos permitirían hacer menor la regresión en el cálculo de las correlaciones con ellas.

La figura muestra la conocida relación en la curva normal entre las puntuaciones típicas, centiles y estaninas.

Según esta técnica preparamos baremos para cada una de las pruebas, con lo que se podía ya calificar a los sujetos y realizar las comparaciones pertinentes.

A continuación presentamos, de modo sintético, las tablas de los datos obtenidos.

Geografía 1.

Tabla de frecuencias y porcentajes acumulados

X	f	fa.	% a.
93-98	8	183	99'99
87-92	11	165	95'37
81-86	16	154	89'01
75-80	19	138	79'76
69-74	22	119	68'78
63-68	24	97	56'07
57-62	18	73	42'19
51-56	11	55	31'79
45-50	18	44	25'43
39-44	15	26	15'03
33-38	3	11	6'36
27-32	2	8	4'62
21-26	5	6	3'47
15-20	1	1	0'58

Baremo

%	St.	PD.
	9	
96	—	93
	8	
89	—	86
	7	
77	—	79
	6	
60	—	70
	5	
40	—	61
	4	
23	—	50
	3	
11	—	42
	2	
4	—	29
	1	

		Error típico	*Error de medida*
$\bar{X}$	= 64'18	1'36	2'66
Mdn.	= 65'87	1'70	3'32
σ	= 17'82	0'96	1'88

Geografía 2.

Tabla de frecuencias y porcentajes acumulados

X	f	fa.	% a.
64-66	1	167	99'99
61-63	17	166	99'40
58-60	7	149	89'22
55-57	17	142	85'03
52-54	32	125	74'85
49-51	30	93	55'69
43-45	14	48	28'74
43'45	14	48	28'74
40-42	8	34	20'36
37-39	15	26	15'57
34-36	4	11	6'58
31-33	3	7	4'19
28-30	3	4	2'39
25-27	1	1	1'60

Baremo

%	St.	PD.
	9	
96	—	63
	8	
89	—	61
	7	
77	—	55
	6	
60	—	52
	5	
40	—	49
	4	
23	—	42
	3	
11	—	38
	2	
4	—	33
	1	

		Error típico	*Error de medida*
$\bar{X}$	= 49'39	0'64	1'21
Mdn.	= 50'55	0'80	1'57
σ	= 8'22	0'45	0'88

Geografía 1 bis.

Tabla de frecuencias y porcentajes acumulados

X	f	fa.	% a.
63-67	2	239	99'99
58-62	19	237	99'16
53-57	45	218	91'21
48-52	43	173	72'38
43-47	48	130	54'39
38-42	35	82	34'30
33-37	19	47	19'66
28-32	17	28	11'71
23-27	9	11	4'60
18-22	0	2	0'83
13-17	2	2	0'83

Baremo

%	St.	PD.
	9	
96	—	67
	8	
89	—	63
	7	
77	—	59
	6	
60	—	52
	5	
40	—	47
	4	
23	—	42
	3	
11	—	36
	2	
4	—	28
	1	

			Error típico	*Error de medida*
$\bar{X}$	=	54'54	0'71	1'39
Mdn.	=	46'40	0'89	1'74
σ	=	9'70	0'50	0'98

Historia A.

Tabla de frecuencias y porcentajes acumulados				*Baremo*		
				%	St.	PD.
X	f	fa.	% a.		9	
				96	—	87
98-103	1	247	99'99		8	
92-97	3	246	99'59	89	—	78
86-91	8	243	98'38		7	
80-85	14	235	95'14	77	—	70
74-79	19	221	89'43		6	
68-73	21	202	81'78	60	—	59
62-67	23	181	73'27		5	
56-61	24	158	63'97	40	—	47
50-55	25	134	54'25		4	
45-49	33	109	44'13	23	—	39
38-43	28	76	30'77		3	
32-37	24	48	19'43	11	—	32
26-31	15	24	9'72		2	
20-25	7	9	3'64	4	—	26
14-19	2	2	0'81		1	

		Error típico	*Error de medida*
$\bar{X}$	= 54'64	1'16	2'27
Mdn.	= 52'98	1'45	2'84
σ	= 18'18	0'82	1'60

Historia B.

Tabla de frecuencias y porcentajes acumulados				*Baremo*		
				%	St.	PD.
X	f	fa.	% a.		9	
				96	—	82
85-89	4	239	99'99		8	
80-84	10	235	98'32	89	—	74
75-79	14	225	94'14		7	
70-74	24	211	88'28	77	—	69
65-69	28	187	78'24		6	
60-64	29	159	66'52	60	—	62
55-59	26	130	54'39		5	
50-54	35	104	43'51	40	—	54
45-49	23	69	28'86		4	
40-44	18	46	19'25	23	—	46
35-39	11	28	11'71		3	
30-34	12	17	7'11	11	—	39
25-29	3	5	2'09		2	
20-24	2	2	0'84	4	—	31
					1	

		Error típico	*Error de medida*
$\bar{X}$ =	57'33	0'92	1'81
Mdn. =	57'50	1'15	2'25
σ =	14'26	0'65	1'27

FIABILIDAD

Solemos decir de los instrumentos de medida que son tanto más fiables cuanto menos sujetos están a alteraciones. La fiabilidad de un instrumento depende pues de su inalterabilidad, y puede comprobarse esto estudiando la estabilidad de los resultados obtenidos con ellos en distintas ocasiones. De una "prueba objetiva" diremos que es tanto más fiable cuanto más constante se mantenga la ordenación de las puntuaciones que un grupo de sujetos obtenga en ella en distintas ocasiones.

De las diversas formas en que es posible calcular la fiabilidad, nos servimos en cada caso de la que nos pareció más conveniente.

a) En las *pruebas de Geografía* se usó el método de las *dos mitades equivalentes,* consistente en hallar la correlación entre las puntuaciones obtenidas en las dos mitades (en este caso ítems pares e impares).

Las correlaciones se calcularon con los datos de la muestra total, por medio de la fórmula de Pearson [5] y se corrigieron los resultados con la fórmula de Spearman Brown [6], dado que este sistema reduce la longitud del test y, como es sabido, esto afecta a la fiabilidad.

b) En las *pruebas de Historia* la fiabilidad se calculó *comparando los resultados en dos pruebas paralelas.* Es el caso que las dos pruebas de Historia, A y B, se pasaron el mismo día a los mismos sujetos, y su contenido podía considerarse equivalente, pues los ítems de los mismos se referían a una misma materia y se habían elegido en proporciones semejantes al contenido del programa de examen. Es decir, que las dos pruebas, en principio, podían considerarse como idénticas, o como partes paralelas de una sola.

A continuación presentamos los índices de fiabilidad obtenidos en cada una de las pruebas, expresando al mismo tiempo el error típico de medida [7] y el error de medida [8] que a un nivel de confianza del 5 por 100 afecta a las puntuaciones.

[5]
$$r_{xy} = \frac{\frac{\sum fx'y'}{N} - \left(\frac{\sum fx'}{N}\right)\left(\frac{\sum fy'}{N}\right)}{\sqrt{\left[\frac{\sum fx'^2}{N} - \left(\frac{\sum fx'}{N}\right)^2\right]\left[\frac{\sum fy'^2}{N} - \left(\frac{\sum fy'}{N}\right)^2\right]}}$$

[6]
$$r_{xx} = \frac{2r_{11}}{1 + r_{11}}$$

[7]
$$\sigma_1 = \sigma_x \sqrt{1 - r_{xx}}$$

[8] E. M. $= \sigma_x \cdot 1'96.$

Pruebas	*Indice de fiabilidad*	*Error típico de medida*	*Error de medida a N. C. de 5 %*
Geografía 1	.92	5'16	10'11
Geografía 2	.87	2'88	5'64
Geografía 1 bis	.97	1'91	3'74
Historia A	.75	9'05	17'73
Historia B	.75	7'10	13'91

Al comparar estos índices de fiabilidad con el de .68 obtenido en el ensayo previo a nuestro trabajo, nos parece haber logrado perfeccionar nuestras pruebas al menos en este aspecto. Sin embargo hay que tener en cuenta que este aumento en parte podría ser explicado por la mayor longitud de las pruebas y no sólo por su perfeccionamiento.

Por medio de la fórmula

$$R_{xx} = \frac{n\ r_{xx}}{1 + (n - 1)\ r_{xx}}$$

en la que R_{xx} indica la fiabilidad que debe tener una prueba cuando se aumenta *n* veces su extensión, partiendo de r_{xx}, la fiabilidad ya conocida, nos sirvió a nosotros para comprobar hasta qué punto el aumento de los índices podía o no explicarse sólo por la mayor longitud de las pruebas usadas.

En el caso de las pruebas de Geografía los resultados fueron claramente favorables, poniendo de manifiesto que habíamos mejorado y mucho en cuanto a la fiabilidad. No sucedió así con el índice de fiabilidad hallado en Historia A y B, cosa acaso debida a la forma en que dicho índice fue calculado.

Otro argumento en favor de la fiabilidad de las pruebas se obtuvo al estudiar la significación de las diferencias entre las puntuaciones de dos grupos de sujetos considerados similares. Es al caso que las pruebas Geografía 1 y Geografía 1 bis contenían una parte totalmente idéntica y se habían administrado a dos cursos diferentes. En principio bien es verdad que los estudiantes de un curso y los de otro pueden ser distintos en su capacidad mental y en su aprovechamiento, pero estas diferencias en cursos normales no suelen ser muy grandes; quiere decirse, que una misma prueba, administrada a sujetos semejantes en condiciones similares, debe dar resultados muy parecidos. Al objeto de verificar esto, se partió de los siguientes datos previamente obtenidos:

	Geog. 1	Geog. 1 bis	dif.
N	173	239	—
$\bar{X}$	14'22	15'06	0'84
σ	4'06	3'38	—

La cuestión era la siguiente: ¿Hasta qué punto estas diferencias podían ser consideradas como producto del azar, esto es, como nulas? Partiendo pues de la hipótesis nula, consideramos si los datos presentados nos permitían rechazarla. Para ello, señalado un nivel de confianza de 5 por 100, pasamos a hallar el error típico de la diferencia entre las medias (σd) según la fórmula

$$\sigma_d = \sigma_{\bar{X}_1} - \sigma_{\bar{X}_2} = \sqrt{\sigma_{\bar{X}_1}^2 + \sigma_{\bar{X}_2}^2} = 0'728 \quad (9)$$

y la razón crítica (Rc), según

$$Rc = \frac{d}{\sigma_d} = 1'15 < 1'96$$

Comoquiera que la razón crítica no es mayor de 1'96, no podemos rechazar la hipótesis nula. Es decir, que podemos afirmar a un nivel de confianza del 5 por 100 que la diferencia entre las medidas no es significativa. Ello claro, es un argumento en favor de la fiabilidad de la prueba, al mismo tiempo que en favor de la similitud de las muestras de sujetos con que se trabajó.

VALIDEZ

De un test podemos decir que es válido si efectivamente mide aquello que se pretende medir. Ahora bien, la dificultad misma que entraña el objeto mensurable, así como la no menor que supone la precisión del instrumento, hace que la validez deba ser expresada en términos de aproximación. Es decir, que la validez se ha de expresar en términos que indiquen hasta qué punto un test cumple con el fin para el cual es destinado.

Así, pues, al preguntarnos hasta qué punto un test es válido, lo primero que hemos de saber con claridad es qué se pretende medir con él. Esto es, que el término validez supone una relación con un fin. Validez, ¿para qué?

Esta validez debe comprobarse empíricamente, por lo que el primer problema al que nos hemos de enfrentar es el de definir, con la mayor exactitud posible, el fin o los fines a que se destina el test. Con ello podremos fijar los criterios de validez.

A) En nuestro caso, en primer lugar, con las "pruebas objetivas", de modo inmediato, lo que se pretende medir es hasta qué punto el alumno ha adquirido un determinado número de conocimientos específicos. En tal

[9] YELA, M.: *Apuntes de Psicometría*, p. 196.

caso nosotros podríamos decir que efectivamente la prueba sería válida en la medida en que aquellos sujetos que obtuvieran puntuaciones más altas en ella, fueran también los que mejor conocieran la materia sobre la que se había preparado el test.

Supongamos que fuera posible preguntar con todo detalle a cada uno de los alumnos, sobre el contenido total de la materia acotada para el examen. Al terminar dicho examen podríamos ordenar a los sujetos según la mayor o menor cantidad de respuestas exactas. Podríamos decir que sabían aquellos que respondían más y con mayor exactitud. Tal concepto de saber es excesivamente cuantitativo, pero de momento puede sernos útil tal consideración.

Ahora bien, nos sería fácil comprobar que la ordenación de los alumnos se mantenía prácticamente idéntica, si se consideraban como preguntas válidas sólo un número determinado de las efectuadas, bien que lo más representativo de todas ellas.

En este concepto muestral es en lo que se basa la *validación del contenido*. Diremos de una "prueba objetiva" que es más o menos válida en la medida en que el contenido de sus ítems sea una muestra más representativa del total de los asertos o puntos fundamentales de una determinada materia. Pero tal representación no debe ser puramente cuantitativa, pues es bien sabido que siempre hay en una materia conocimientos más importantes que otros. Así, pues, la selección de los ítems no se habrá de hacer simplemente al azar, sino por un muestreo estratificado de acuerdo con la importancia relativa de los diversos puntos de la materia de examen; esto es, de acuerdo con una previa valoración cualitativa a juicio de los profesores de la referida materia.

En nuestro caso esto se mostró en tablas de especificación en las que, junto a una relación minuciosa de los capítulos de la materia de que se trataba, se hizo ver de dónde y en qué proporción se habían elegido los ítems de las pruebas.

B) Por otro lado, si lo que se pretende medir con las pruebas es hasta qué punto los sujetos han alcanzado un cierto nivel de instrucción, debería de suceder que los resultados del test coincidieran con las estimaciones que por otros medios se hicieran de dicho nivel de instrucción o aprovechamiento. Tal estimación no puede ser considerada como absolutamente fiable, pero en la medida en que los resultados de las pruebas y dichas estimaciones coincidan, podremos decir que las pruebas sirven para medir, al menos en parte, lo que tales estimaciones tengan de varianza no aleatoria.

Supongamos que al tiempo de pasar las "pruebas objetivas", hubiéramos sometido a los alumnos a otro tipo de examen. Podríamos hablar de validez en la medida en que las puntuaciones de ambos tipos de exámenes coincidieran.

Pues bien, en nuestra tesis hemos seguido asimismo esta segunda vía. En las "pruebas objetivas" se incluyó a veces un tipo de cuestiones (parte 4.ª de las pruebas, Cuestiones breves) que en principio podían ser consideradas como un examen paralelo de tipo cuasi-tradicional y por tanto utilizable como criterio simultáneo e interno de la prueba misma.

Podremos, pues, hablar de *validez concurrente o simultánea,* en la medida en que las puntuaciones, alcanzadas por los alumnos de esta parte cuarta de la prueba, sean más o menos similares a las que obtienen en el total de las tres primeras partes de la prueba.

Así se estudió en las pruebas de Geografía 1 y en las de Historia A y B, en las que se obtuvo los siguientes índices de validez:

Pruebas	r_{xy}
Geografía 1.	.79
Historia A.	.63
Historia B.	.73

Ahora bien, así como usamos este tipo de cuestiones breves como criterio, igualmente podríamos considerar otras que fueran de la misma significación. Se comprenderá, pues, que la fiabilidad de las pruebas, así como la intercorrelación de sus partes pueden ser también usadas como índice de la validez concurrente.

C) Finalmente, los resultados de las pruebas se pueden considerar como fundamento de un pronóstico aproximado del éxito o fracaso académico del alumno en un futuro próximo. Es decir, que si en un momento determinado el sujeto ha sido capaz de resolver una serie de problemas concretos —"prueba objetiva"—, lo cual supone una capacitación específica, cabe esperar, supuesta su dedicación y entrega al estudio, que en el futuro será capaz de superar otras pruebas similares —exámenes tradicionales—.

Podríamos decir que nuestras pruebas tienen una determinada *validez predictiva,* en la medida en que sus resultados sean más o menos similares a los obtenidos en dichos exámenes.

Tales predicciones no podrán ser muy exactas, en primer lugar, porque durante el tiempo que transcurre desde la aplicación de la prueba hasta el examen final de curso pueden intervenir factores que modifiquen diferencialmente las condiciones del alumno, y en segundo, porque la fiabilidad de los exámenes tradicionales suele ser bastante baja.

En las pruebas de Geografía, el criterio externo de que nos servimos para su validación predictiva fue el de las calificaciones obtenidas por los alumnos en el examen de final de curso. Tal examen se realizó según la forma tradicional, siendo invitados los alumnos a responder por escrito a tres temas.

La distribución de las calificaciones fue la siguiente:

No presentados	= 21
0	= 2
1	= 2
2	= 2
3	= 8
4	= 29
5	= 36
6	= 30
7	= 19
8	= 11
9	= 2
10	= 0
Total	=162

No pudimos estudiar la fiabilidad de dichas calificaciones, debido a que los resultados se nos presentaron en conjunto y no como notas por separado en cada uno de los temas del examen.

La distribución de estas calificaciones nos pareció que podía ser considerada como normal, y por ello, tomando estos resultados como criterio, bien que prescindiendo de los sujetos no presentados, procedimos a hallar la correlación de Pearson entre las puntuaciones estaninas obtenidas por los sujetos en cada una de las pruebas y las referidas calificaciones alcanzadas en el examen final. Los índices de validez obtenidos fueron los siguientes:

Pruebas	r_{xy}
Geografía 1.	.53
Geografía 2.	.75

Para que pueda verse mejor en qué consiste este valor predictivo vamos a presentar en sendas tablas de expectancia la comparación entre los resultados de las pruebas y las calificaciones del examen final.

Geografía 1.

*Puntuaciones en la "prueba objetiva"**	*Calificaciones en el examen final*			
	No presentados y suspensos	*Aprobados*	*Notables y sobresalientes*	*Total*
Superiores	4	15	20	39
Promedio	34	41	12	87
Inferiores	26	10	0	36
Total	64	66	32	162

Geografía 2.

*Puntuaciones en la "prueba objetiva"**	*Calificaciones en el examen final*			
	No presentados y suspensos	*Aprobados*	*Notables y sobresalientes*	*Total*
Superiores	3	16	19	38
Promedio	32	39	13	84
Inferiores	29	11	0	40
Total	64	66	32	162

Se advertirá inmediatamente la gran superioridad del valor predictivo de la prueba de Geografía 2. en comparación con la Geografía 1. Dado que el contenido de las pruebas y el del examen final era diferente, hemos de pensar que tal superioridad es debida a su mayor proximidad en el tiempo respecto al examen final. Recuérdese que la prueba de Geografía 1. se pasó en enero y la Geografía 2. en abril, mientras que los exámenes finales tuvieron lugar en junio. Esta mayor coincidencia entre los sujetos que contestan mejor a la prueba de Geografía 2. y al examen final, en comparación con los que lo hacen mejor a la de Geografía 1. y a dicho examen final, creemos que es debida a que en los momentos más avanzados del curso, la capacitación de los sujetos en una materia está mucho más definida y, en general, los que contestan bien a un tipo de prueba lo hacen también a la otra. Creemos, pues, poder concluir que la capacidad

* Se consideraron superiores los estaninos 7, 8 y 9, equivalentes a 23 %; promedio, a los estaninos 4, 5 y 6, equivalentes a 54 %, e inferiores a los estaninos 1, 2 y 3, equivalentes a 23 %.

predictiva de las pruebas aumenta en la medida en que el tiempo entre ellas y el examen final es menor, por lo que pensamos que, si se pasaran al final del curso y sobre materia idéntica, los resultados serían todavía más semejantes, y en consecuencia, se podría prescindir del examen de tipo tradicional.

En las pruebas de Historia el criterio externo de que nos servimos fue también el de las calificaciones obtenidas por los alumnos en el examen de final de curso. En tal examen los sujetos fueron examinados por grupos, en distintos días, por el método tradicional de exámenes escritos, sobre diversos temas, u oralmente en aquellos casos en que su examen escrito ofrecía dudas para la calificación. El criterio, pues, como podrá advertirse, no se prestaba a un estudio de su fiabilidad.

Los resultados para los sujetos que habían pasado las "pruebas objetivas" en enero eran los siguientes:

No presentados =	35
Suspensos =	82
Aprobados =	76
Notables =	22
Sobresalientes =	3
Total =	218

Dado que la variable de las calificaciones no se nos presentó en forma continua, para comparar éstas con las puntuaciones alcanzadas por los sujetos en la "prueba objetiva", recurrimos a hallar correlaciones tetracóricas entre las variables. Dichas variables se dicotomizaron del modo siguiente: Variable de calificaciones, suspensos y no presentados en el grupo inferior, aprobados y notas en el grupo superior. Variable de puntuaciones estaninas, del 1 al 4, inclusive, en el grupo inferior y del 5 al 9 en el superior.

Los índices de correlación obtenidos fueron los siguientes:

Pruebas	r_{xy}
Historia A.	.71
Historia B	.60

Se advertirá en seguida una cierta superioridad de la prueba de Historia A sobre la de Historia B. Pero, ¿cómo podía explicarse esto si ambas pruebas habían sido preparadas con idénticos criterios y el contenido de los ítems podía decirse igualmente representativo de la materia acotada para el examen? Las pruebas, además, se habían pasado al mismo tiempo y su amplitud era prácticamente la misma. ¿Por qué, pues, una de ellas, al parecer, pronosticaba mejor el criterio?

Se nos ocurrió entonces otra pregunta: ¿Realmente era así? ¿Se podía decir, por los índices hallados, que efectivamente una prueba pronosticaba mejor que la otra? En otras palabras, ¿eran significativas las diferencias de estos índices de validez?

Para resolver esto recurrimos a estudiar la significación de las diferencias entre los índices de correlación hallados [10].

Señalando un nivel de confianza del 5 por 100, recurrimos a las tablas de Fisher para convertir los valores r en z. Después, mediante la fórmula

$$Rc = \frac{|z_1 - z_2|}{\sigma_{z_1 - z_2}} = \frac{d}{\sigma_d}$$

en la que $\sigma_d = \sqrt{\sigma_{z_2}^2 + \sigma_{z_2}^2}$

y, sabiendo que $\sigma_{z_1}^2 = \frac{1}{\sqrt{N_1 - 3}}$ y $\sigma_{z_2}^2 = \frac{2}{\sqrt{N_2 - 3}}$

siendo N_1 y $N_2 = 218$,

pudimos realizar nuestros cálculos, llegando a los siguientes resultados:

zH_A [11]	zH_B [12]	d	σ_d	d/σ_d
.887182	.693146	.174153	.3693	.5254156 < 1'96

Donde se ve que no es significativa la diferencia entre los índices de validez hallados y que, por consiguiente, dicha diferencia es explicable por azar y no necesariamente por razones internas de las pruebas.

Al objeto de que de un modo más intuitivo pueda compararse esta capacidad predictiva de las pruebas de Historia A y B, vamos a presentar en sendas tablas de expectancia los resultados obtenidos.

HISTORIA A

Puntuaciones en la "prueba objetiva"	*Calificaciones en el examen final*			
	No presentados y suspensos	*Aprobados*	*Notables y sobresalientes*	*Total*
Superiores	12	23	14	49
Promedio	63	51	10	124
Inferiores	42	2	1	45
Total	117	76	25	218

[10] YELA, M.: *Apuntes de Psicometría*, p. 212.

[11] z de Fisher para el índice de validez en Historia A.

[12] z de Fisher para el índice de validez en Historia B.

HISTORIA B

Puntuaciones en la "prueba objetiva"	*Calificaciones en el examen final*			
	No presentados y suspensos	*Aprobados*	*Notables y sobresalientes*	*Total*
Superiores	11	27	16	54
Promedio	59	42	8	109
Inferiores	47	7	1	55
Total	117	76	25	218

Pese a que los índices de validez obtenidos y las mismas tablas de expectancia, realmente nos muestran una cierta inferioridad de los resultados obtenidos con las pruebas de Historia en comparación con los conseguidos en Geografía, no por ello son despreciables. Si se tiene en cuenta además, que estas pruebas se pasaron a los sujetos en enero, con tanta más razón estaremos dispuestos a estimar los referidos valores predictivos.

Conclusiones

Siguiendo todo este largo razonamiento creemos poder concluír:

1.° Que las "pruebas objetivas" pueden utilizarse con toda tranquilidad como sustitutivas de las técnicas de examen tradicionalmente usadas. Esto no quiere decir, claro, que nosotros recomendemos una drástica supresión de las mismas.

2.° Que los exámenes por medio de "pruebas objetivas" ofrecen una serie de ventajas nada despreciables tanto para los profesores —economía de tiempo y fatiga—, como para los alumnos mayor justeza en la calificación.

3.° Creemos además que estos resultados y conclusiones pueden ser aplicados con fruto a otras experiencias en diferentes materias de estudio.

Seminario de Psicología
Valencia

BIBLIOGRAFIA SELECTA

ADKINS, D. C.: "Measurement in Relation to Educational Process". *Educational and Psychological Measurement*, vol. II, pp. 221-240.

AHMANN, S. J.: *Testing Student Achievements and Aptitudes.* Washington: Center for Appl. Res. in Educat. 1962.

BEAN, K. L.: *Construction of educational and personnel tests.* N. Y.: McGraw-Hill. 1956.

BLOOM, B. S. (Ed.): *Taxonomy of Educational Objectives.* N. Y.: David McKay Co. 1956.

DANIELS, J. C.: "Testing Geography at the Ordinary Level of the General Certificate of Education". *The British Journal of Educational Psychology*, vol. XXIV, part. III. Nov. 1954, pp. 180-89. London.

DAVIS, F. B.: "Item Analysis in Relation to Educational and Psychological Testing". *Psych. Bull.* 49, Mar. 1952, pp. 97-121.

DAVIS, F. B.; SCHWAB, J. J.; CARROLL, J. B., y GULLIKSEN, H.: "Criteria for the evaluation of achievement tests". Proc. 1950 invit. *Conf. test. Probl.*; Educ. Test. Serv. 1951, pp. 73-112.

FERNÁNDEZ HUERTAS, J.: "Métodos de consistencia y equivalencia en la determinación de la fiabilidad de las pruebas instructivas". *Rev. Esp. de Ped.* 12 (1954), pp. 421-27.

FREEMAN, F. S.: *Theory and Practice of Psychological Testing.* N. Y.: Holt. 1963.

GREEN, E. E.; JORGENSEN, A. N., y GERBERICH, J. R.: *Measurement and evaluation in the Secondary School.* N. Y.: 1947.

GULLIKSEN, H.: *Theory of Mental Tests.* N. J.: Wiley, 1962.

HAWKES, H. E.; LINDQUIST, E. F., and MANN, C. R. (Eds.): *The construction and use of achievement examination.* Boston: Hougton Mifflin Co. 1936.

HORST, A. P.: "A general Expression for the Reliability of Measures". *Psychometrica*, 14 (1949), pp. 21-32.

HOTYAT, F.: *Les examens. Les moyens d'evaluation dans l'enseignement.* París: Ed. Bourrelier, 1962.

LINDQUIST, E. F.: "Preliminary Considerations in Objective Test Construction". *Educational Measurement.* Washington: Amer. Counc. on Educat. 1951.

STANLEY, J. C., and DALE, L. B.: "Taxonomy of Educational Objectives". *Educational and Psychological Measurements.* 17 (1957), pp. 632-33.

WOOD, D. A.: *Test Construction. Development and Interpretation of Achievement Tests.* Marrill. Ohio, 1964.

WOODY, C.: "Measurements of some Achievements in Arithmetic School and Society". 4 (1916), pp. 299-303. Teachers College. Columbia University Contribution to Education. No. 80, 1916.

YELA, M.: *Apuntes de Psicometría.* Madrid: Esc. de Psicol. 1960.

YELA, M., y PASCUAL, M.: "El test como instrumento científico". *Rev. de Psicol. Gral. y Aplic.*, vol. XXIX, núm. 74. Madrid, 1964.