Las pruebas objetivas: normas, modalidades y cuestiones discutidas

• Universidad Pontificia Comillas • Madrid •
Facultad de Ciencias Humanas y Sociales

Índice

1. Introducción

Al tratar de las *normas, modalidades y cuestiones discutidas* en las pruebas objetivas, partimos de la constatación de estas realidades:

1. Las pruebas objetivas *(tipo test)* son muy utilizadas, debido sobre todo a la *facilidad de corrección* comparadas con las pruebas tradicionales de respuesta abierta. Esta ventaja es más apreciable si se dispone de correctora de *lectura óptica* y cuando los alumnos son muchos[1].

2. Los profesores en general no están preparados para hacer buenas preguntas objetivas; fácilmente se redactan más preguntas de lo que es más fácil preguntar (y no de lo que se debería preguntar), hay un número excesivo de preguntas memorísticas y descontextualizadas que condicionan en el alumno un modo de estudiar muy pobre, frecuentemente no hay coherencia entre objetivos y evaluación, etc.; es decir, abundan las pruebas objetivas de mala calidad con consecuencias negativas para la moral de los alumnos y la calidad de su aprendizaje[2].

3. El profesor toma sus decisiones (sobre cómo redactar las preguntas, cuántas respuestas poner, aplicar o no aplicar la *fórmula de corrección por adivinación*, etc.) guiado por su sentido común o imitando lo que hacen otros, pero ni el sentido común ni la imitación de prácticas muy generalizadas son en estos temas una garantía de estar haciendo lo más correcto.

4. Finalmente sobre estos temas relacionados con las pruebas objetivas existe un cúmulo grande de investigaciones experimentales, que pueden servir de orientación para mejorar la calidad de estas pruebas y que la mayoría de los profesores suele desconocer.

Aquí vamos a analizar una serie de problemas y cuestiones más o menos discutidas en relación con estas pruebas desde una perspectiva casi puramente experimental: no se trata fundamentalmente de exponer lo que otros *piensan* sino *lo que sucede de hecho*, tal como se desprende de la abundante investigación sobre las pruebas *tipo-test.* A veces se puede llegar a conclusiones muy claras (por ejemplo sobre el número óptimo de respuestas) y otras veces no,

[1] Hay que recordar que también va habiendo programas de ordenador para corregir preguntas abiertas; Shermis y otros (2002) analizan dos experimentos, en uno se corrigen 807 ensayos y en otro 386; además de corregirlos con un programa de ordenador son corregidos por 6 correctores; la correlación entre correctores es de .71 y la correlación entre los correctores y el ordenador es de .83

[2] Paxton (2000) presenta una buena crítica al abuso de pruebas objetivas de mala calidad en la Universidad y a sus consecuencias en el aprendizaje de los alumnos.

pero en todos los casos se ofrecen conclusiones de muchos estudios experimentales que pueden orientar a cualquier profesor a tomar sus propias decisiones y además ofreciendo una justificación.

Como guía orientadora, podemos ver lo que da de sí la investigación en torno a estas cuestiones:

- Son muchas las *normas* que suelen darse para redactar bien estas preguntas ¿Son importantes y para qué son importantes estas normas?

- A mayor número de respuestas será más difícil adivinar la respuesta correcta sin conocerla, pero ¿Existe alguna recomendación fundada en datos experimentales sobre el *número óptimo* de respuestas?

- Las preguntas del tipo *verdadero-falso* son muy habituales y aparentemente fáciles de formular: ¿Cuáles son sus problemas? ¿Cómo se pueden mejorar?

- Lo habitual es que haya una *única respuesta correcta:* ¿qué puede decirse de las preguntas con varias respuestas correctas?

- El problema clásico de estas pruebas objetivas es que el que responde puede *adivinar la respuesta* sin conocerla: ¿Cómo se puede *valorar* la fórmula de corrección por adivinación? ¿Mejoran las características psicométricas de un test aplicando esta fórmula? ¿Puede una respuesta correcta aparentemente adivinada ser fruto de un conocimiento inseguro y parcial?

- ¿Hay otros modos de corregir estas pruebas que controlen en cierto grado la adivinación sin necesidad de aplicar fórmulas correctoras?

Vamos a tratar por lo tanto seis temas relacionados con las pruebas objetivas. La extensión dada a estos temas es muy desigual; en todos los casos se trata de cuestiones en las que hay una abundante investigación experimental y que con frecuencia se prestan a discusión. Los temas tratados son los siguientes:

1º Cuestiones relacionadas con las *normas* que suelen darse para la redacción de las preguntas objetivas y sus respuestas
2º El *número de respuestas* en los ítems *(¿existe un número óptimo?)*
3º Las preguntas del tipo *Verdadero-Falso*
4º Las preguntas con *dos alternativas* distintas del *Verdadero-Falso*

5º Las preguntas con *varias respuestas correctas:* diversas maneras de presentar y corregir estas preguntas.

6º La fórmula de *corrección por adivinación* y otras alternativas.

Un punto que conviene destacar es que *el tipo de pregunta o examen esperado condiciona el cómo estudia el alumno,* y de cómo estudia el alumno va a depender en buena medida el cómo se forma, cómo aprende a pensar, etc. No se trata ya de preparar pruebas objetivas sin errores, sino de preparar preguntas que estimulen en el alumno un estudio inteligente. No vamos a tratar o proponer aquí directamente modelos de *preguntas inteligentes* que pueden encontrarse en otras publicaciones[3], pero sí es oportuno recordar que con demasiada frecuencia las pruebas objetivas, aunque estén formalmente bien hechas, estimulan en el alumno un estudio muy pobre[4]. En términos generales y para estimular en el alumno un estudio menos memorístico y más rico, se pueden hacer dos recomendaciones:

1ª No limitarse a las pruebas objetivas como método de evaluación casi exclusivo porque aunque estén muy bien hechas dejan fuera aspectos importantes del estudio y de su evaluación, como es todo lo relacionado con la expresión escrita; las pruebas objetivas deben formar parte de un espectro más amplio de métodos de evaluación, como suele indicarse con frecuencia al tratar de métodos de evaluación (como Paxton, 2000).

2ª Si se va a utilizar este tipo de preguntas, conviene examinar buenos modelos de preguntas objetivas, porque es un error pensar que las preguntas objetivas se limitan necesariamente a comprobar conocimientos memorísticos.

2. Normas en la redacción de las pruebas objetivas

Al redactar preguntas objetivas es fácil cometer errores, tanto en la misma formulación de la pregunta como en la redacción de las respuestas. Sobre el *arte* de redactar pruebas objetivas hay una abundante literatura y casi todos los libros de evaluación dan una serie de normas sobre cómo hacerlo[5]. En parte se trata de normas que podemos considerar de sentido

[3] Un excelente texto de evaluación con buenos modelos de preguntas de todo tipo es el de Bloom, Madaus y Hastings (1981), de título muy expresivo: *evaluation to improve learning*.

[4] Una investigación típica sobre cómo el examen esperado condiciona el *cómo* estudia el alumno es de Scouller (1998, con 206 alumnos de Educación). Cuando el alumno espera una prueba objetiva, el estudio es más superficial y memorístico, y cuando espera preguntas abiertas el estudio es más profundo; además un enfoque superficial en el estudio está asociado a buenos resultados en un examen *tipo test,* y a peores resultados en un examen de preguntas abiertas; y al revés un enfoque profundo está asociado a buenos resultados en la prueba abierta y a peores en la prueba objetiva. Esta cuestión no es sin embargo tan simple porque hay muchos tipos de preguntas objetivas, pero sí es verdad que las pruebas objetivas más habituales y fáciles de preparar (para el profesor) condicionan un estudio más bien superficial.

[5] Normas y ejemplos para redactar preguntas objetivas (y también sobre otros sistemas de evaluación) pueden encontrarse con facilidad en textos de evaluación (uno excelente es el de Bloom, Madaus y Hastings, 1981) y en Internet, por ejemplo Burton, Sudweeks, Merrill y Wood (1990, del Testing Center de Brigham Young University), Gross (1980) y en las páginas Web de muchas universidades (por ejemplo University of Minnesota, The Office of Measurement Services, en Classroom Resources).

común (como evitar que la respuesta correcta sea de una longitud desproporcionada comparada con las respuestas falsas), pero en la práctica, debido a la prisa, poco cuidado o no caer en la cuenta de los errores que se pueden cometer, es frecuente ver preguntas redactadas en contra de las normas usuales.

Con estas normas se pretende:

a) que las preguntas sean claras; la tarea del alumno es pensar y escoger la respuesta correcta, no adivinar qué se le está preguntando;

b) que en la misma redacción de la pregunta o de las respuestas no se den pistas sobre cuál es la respuesta correcta, tanto de la misma pregunta como de otras dentro del mismo test;

c) que las preguntas no favorezcan o perjudiquen a determinados tipos de alumnos, independientemente de sus conocimientos.

Estas normas están publicadas en numerosos textos; Haladayna y Downing (1989) han revisado 46 textos de evaluación que vienen a coincidir en 43 normas para redactar buenas preguntas objetivas[6]. Pueden parecer demasiadas; unas son de sentido común (sentido común olvidado con frecuencia); otras parecen menos obvias pero tienen un fundamento más claro en la ya amplia investigación sobre estos temas.

Las críticas a las pruebas objetivas son frecuentes y podemos afirmar con Armstrong (1993) que muchas de estas críticas no se derivan necesariamente del formato mismo de las pruebas objetivas, sino de la mala calidad presente en muchos tests. Son muchos los que tienen una experiencia personal muy negativa por haber tenido que responder a pruebas objetivas mal hechas.[7]

La calidad de una prueba objetiva no es de igual importancia en todas las situaciones. No es lo mismo la calidad de un sencillo examen parcial cuyos resultados van a ser de importancia menor o se van a complementar con otros muchos datos de los alumnos, que un examen de selección o admisión de cuyo resultado se van a derivar consecuencias muy importantes para el alumno. En cualquier caso es importante siempre hacer estas pruebas lo mejor posible.

---

[6] Una breve exposición de las normas más importantes puede verse en Frary (1995) y Kehoe (1995).

[7] Como el mismo Armstrong reconoce (1993) estas críticas son menos aplicables a los tests publicados, en los que suelen seguirse con más rigor las normas establecidas para la construcción de tests y que por lo tanto tienden a ser de una calidad superior; el problema de la calidad es más frecuente en los tests que hace cada profesor cuando prepara sus propios exámenes. El autor utiliza en su estudio alumnos universitarios acostumbrados a responder a este tipo de tests.

En este apartado no nos vamos a extender en las normas que suelen darse para hacer buenas preguntas objetivas; sólo nos vamos a fijar en algunos puntos relacionados con este tema que tienen apoyo claro en estudios experimentales; de todas maneras en el primer punto tratado ya aparecen enunciadas las normas más frecuentes.

2.1. Normas para evitar pistas que faciliten responder correctamente sin conocer la respuesta

Sobre el adivinar en las preguntas objetivas y la fórmula de *corrección por adivinación* tratamos extensamente más adelante. Ahora nos fijamos en las *normas* para redactar estas preguntas y que tienen por objeto precisamente no dar facilidades para responder correctamente sin conocer la respuesta.

En las preguntas mal formuladas suele haber *pistas* sobre cuál es la respuesta correcta o al menos para descartar algunas alternativas incorrectas, Hay alumnos que tienen una habilidad especial para descubrir estas pistas; es más esta habilidad puede ser enseñada y aprendida incluso por niños.

Esta habilidad específica (responde bien sin conocimientos suficientes) tiene su propio y expresivo nombre en inglés, *test-wiseness*. Esta habilidad consiste en *saber ver* (o intuir) qué alternativas son probablemente correctas o incorrectas, o más genéricamente, *test-wiseness* es la capacidad de saber utilizar las características y formato del test (e incluso de la situación) para mejorar los resultados; esta *habilidad* es independiente de los conocimientos que pueda tener el que responde al test[8].

Morse (1998) analiza la eficacia de las distintas normas para no dar pistas indebidas; presenta los resultados de siete investigaciones hechas con niños[9] y de cuatro hechas con alumnos universitarios[10]; no todas las normas son igualmente eficaces con todas las edades.

Con niños las normas que mejor funcionan son (por este orden):

1) *Evitar alternativas absurdas,*
2) *Evitar el parecido entre la pregunta y alguna de las alternativas* (suele ser la respuesta correcta),

[8] Una exposición clara de esta habilidad *(test-wiseness)* y una taxonomía de *estrategias* empleadas por los alumnos puede verse en Rogers y Yang (1996).

[9] Una de estas investigaciones hechas con niños, y con ejemplos de preguntas que no siguen las normas habituales, es la de Carter (1986).

[10] La última de estas investigaciones está hecha por el mismo autor; utiliza una muestra de 243 alumnos de tres universidades distintas; el instrumento utilizado está diseñado para medir esta habilidad *(test-wiseness).*

3) *Evitar alternativas muy semejantes* (si sólo hay una respuesta correcta, serán falsas las dos),
4) *Evitar el uso de determinantes muy específicos* (como *siempre, nunca,* etc.).

Con universitarios las normas en conjunto más eficaces son estas cuatro:

1) *Evitar falta de concordancia gramatical*
2) *Evitar que la respuesta correcta sea más larga, más elaborada*
3) *Evitar alternativas absurdas*
4) *Evitar el uso de determinantes muy específicos* (como *siempre, nunca,* etc.).

Es claro que no seguir este tipo de normas puede hacer un test artificialmente más fácil de lo pretendido.

Rogers y Harley (1999) también muestran que cuando se utiliza un número mayor de alternativas (cuatro en vez de tres), aumentan también las pistas indebidas para rechazar alternativas falsas[11]; estas investigaciones son coherentes con las que muestran que cuando se eliminan las alternativas menos funcionales (no atraen ni a los que menos saben), aumenta la fiabilidad de todo el test (Cizek, Robinson y O'Day, 1998).

2.2. Las pruebas objetivas mal hechas perjudican a los alumnos *dependientes de campo*

En cualquier tipo de examen se pretende comprobar qué sabe el alumno, diferenciar a unos alumnos de otros según sepan más o menos, pero no se busca diferenciarlos según criterios ajenos a su propia competencia en cuanto alumnos, como pueden ser determinadas características de personalidad.

Uno de los puntos estudiados en relación con otras preguntas y sus normas es la relación entre *estilo cognitivo* del alumno y su rendimiento en pruebas objetivas. Por estilo cognitivo entendemos aquí la distinción que hacen (y miden Witkin y sus colaboradores, 1971) entre *dependientes* e *independientes de campo*.

Sin entrar a fondo en la explicación de esta tipología, pero para aclarar suficientemente estos conceptos, podemos decir que los sujetos *independientes* de campo son *más analíticos* y los *dependientes* de campo dependen más del contexto, tienden a ver *totalidades*. Esta diferenciación entre *independientes* y *dependientes* de campo (dependen más o menos del

[11] En estos autores pueden encontrarse más referencias sobre esta habilidad de *test-wiseness;* el primer autor tiene otros estudios sobre lo mismo.

contexto en su percepción de las cosas) es ajena a la inteligencia del alumno y a la calidad de su estudio.

En el caso de las preguntas objetivas, el *campo* es la misma pregunta y sus respuestas. Cuando en una pregunta hay pistas irrelevantes (respuesta correcta más larga, falta de concordancia gramatical entre pregunta y respuestas, etc.) estas *pistas* las perciben mejor los que tienen un estilo cognitivo independiente de campo, que suelen tener mejores resultados en los exámenes, y los dependientes de campo quedan perjudicados. Cuando las preguntas están bien redactadas, siguiendo las normas usuales, esta diferencia en éxito entre independientes y dependientes de campo, desaparece. Esta interacción entre estilo cognitivo y redacción de los ítems está bien confirmada en numerosos estudios (Armstrong, 1993, que además cita muchas otras investigaciones).[12]

A propósito del *estilo cognitivo*, Friedman y Cook (1994) exploran las posibles relaciones entre *dependencia-independencia de campo* (también medida con el test de Witkin) y *cambio en la respuesta*, pues con frecuencia los alumnos repasan el examen, lo piensan mejor, y cambian su primera respuesta. El tema de los *cambios de respuesta* también tiene su interés específico. Estos autores, con una muestra de N = 200, hallan correlaciones negativas y significativas (-.19 y -.32) entre respuestas correctas a dos tests y número de cambios en las respuestas (cambian más los que van peor) y una ligera correlación positiva (.11 y .13) entre efectos del cambio y total en los dos tests (predominan ligeramente los cambios de mal a bien sobre los cambios de bien a mal), y prácticamente no hay ninguna relación entre número de cambios y estilo cognitivo.

La única conclusión generalizable de los estudios vistos es que las preguntas mal redactadas perjudican a los dependientes de campo.

Por lo que respecta al *cambio de respuesta* (que también podría estar inducido por una mala redacción de la pregunta) Friedman y Cook (1994) mencionan numerosos estudios y presentan una bibliografía muy completa sobre este punto. La conclusión más generalizable (en alumnos universitarios) es que son los mejores alumnos quienes suelen beneficiarse del cambio de respuesta (cambian preferentemente de mal a bien). Aunque el cambio de respuesta tiene aquí una importancia menor, Hassmen y Hunt (1994) citan numerosos estudios que

[12] El estudio de Armstrong (1993) confirma que cuando se redactan los ítems siguiendo las normas usuales, no hay diferencias entre los dependientes e independientes de campo. Utiliza en su investigación el *Group Embedded Figures Test* y dos versiones del mismo test de conocimientos, una versión siguiendo las normas de redacción que pueden encontrarse en los manuales de evaluación, y otra dando pistas irrelevantes bien disimuladas en la redacción de los ítems; de hecho los sujetos fueron después incapaces de localizar conscientemente la mayoría de estas pistas.

muestran que en general las mujeres cambian más sus respuestas que los varones, y también tienden a omitir más ítems.

2.3. Sobre la redacción de las preguntas ¿Preguntas propiamente dichas o frases incompletas?

La pregunta *completa* tiene dos componentes, la parte inicial, una pregunta propiamente dicha o una frase incompleta, y las alternativas de respuesta. Cuando la pregunta se formula en forma de frase incompleta, las alternativas de respuesta completan, con concordancia gramatical y sintáctica, el encabezamiento del ítem. En cualquier caso el conjunto debe equivaler a una pregunta en sentido propio, de manera que las respuestas no equivalgan a una serie de afirmaciones inconexas.

En algunos estudios experimentales se comparan los dos modos de formular la pregunta, en forma de pregunta propiamente dicha, entre interrogaciones, o en forma de frase incompleta. Estas investigaciones muestran que cuando se formulan los ítems *en forma de pregunta* en sentido propio:

1º Las preguntas son ligeramente más fáciles que si se trata de frases incompletas,

2º No tiene especial efecto en la discriminación; los ítems discriminan de manera parecida (diferencian a los que más saben de los que menos saben) tanto si se formulan como preguntas como si se presentan como frases incompletas;

3º Suben algo la fiabilidad (en torno a .065) y la validez (correlaciones con otros criterios) de todo el test.

Estas ventajas de la formulación del ítem en forma de pregunta no son siempre importantes ni las confirman todos los estudios (revisión de estudios de Crehan, 1989). Se trata, a la vista de estos análisis, de una cuestión que podemos considerar como menor y sin especial trascendencia para la calidad de las preguntas.

2.4. Sobre el uso de partículas negativas *(no, nunca, excepto, menos)* en la formulación de las preguntas

Los ítems con formulación negativa suelen ser de hecho más difíciles, y esto parece confirmado con alumnos de enseñanza primaria y secundaria[13]. Con alumnos universitarios no

[13] Por analogía con este tema, inclusión de partículas negativas en los ítems, podemos recordar que los estudios sobre instrumentos de medición psicológica (como escalas de actitudes y otros) también muestran que la inclusión de estas partículas hacen bajar la calidad de los instrumentos, sobre todo con niños; en cambio los ítems negativos pero sin

está tan claro; al menos cuando se trata de alumnos bien preparados y acostumbrados a este tipo de pruebas. En este caso les afectan menos las deficiencias en la formulación de los ítems (Downing, 1991; Downing, Grosso y Norcini, 1994).[14]

Como norma general no se deben incluir en la formulación del ítem partículas negativas porque se prestan a equivocaciones aun conociendo la respuesta. Si se utilizan lo que suele recomendarse es que la partícula negativa quede suficientemente destacada (subrayada, en *cursiva* o en MAYÚSCULAS). Las preguntas negativas no son necesariamente malas preguntas; la relevancia de lo que se pregunta, en determinados contextos, puede pedir una formulación negativa (por ejemplo saber excluir síntomas falsos de una enfermedad).

2.5. Información ofrecida en los ítems

Una de las normas que suelen darse para redactar preguntas objetivas, consiste en no introducir en la pregunta datos o elementos que no son necesarios para la respuesta; no dar más información que la que es necesaria (evitar el *window dressing* como se dice familiarmente en inglés) y mantener la información previa lo más reducida posible. En el análisis que hacen Haladyna y Downing (1989) de 46 textos de evaluación, 20 hacen esta recomendación.

Gray y Rachor (1995) investigan las consecuencias de la longitud de las preguntas en exámenes de medicina, con ítems en los que se presentan casos clínicos *(clinically focused ítems).* En total analizan 900 ítems y no detectan ninguna relación apreciable entre longitud del ítem y dificultad; los más largos tienden a discriminar mejor, pero esta relación es muy pequeña. Hay que tener en cuenta que en este estudio los examinados son de nivel alto y localizan con facilidad la información relevante para responder correctamente. Aún así, es preferible mantener las preguntas cortas y no aportar más información de la necesaria por razones al menos de eficiencia (mayor número de ítems respondidos en el mismo tiempo).

2.6. Sobre el uso de *todas las anteriores* y *ninguna de las anteriores* como posibles respuestas

Cuando al que redacta no se le ocurren alternativas válidas como respuesta, y sobre todo si desea, como es aconsejable, mantener un idéntico número de respuestas en todas las

---

partículas negativas en su redacción (el estar *de acuerdo* con la afirmación supone estar *en contra* de lo que se mide, como *estudiar es aburrido* midiendo *actitud hacia el estudio*) suelen ser más discriminantes; de la misma manera que en pruebas de conocimientos del tipo *verdadero-falso* suelen ser más discriminantes los ítems cuya respuesta correcta es *falso* (Morales, 2006, tratando de la *aquiescencia* o tendencia a mostrar acuerdo en caso de duda, en los instrumentos de medición psicológica).

[14] Estos autores analizan dos exámenes de medicina (para obtener la licencia para ejercer); en uno se analizan 879 ítems respondidos por 11.454 alumnos; en otro analizan 802 ítems respondidos por 7.178 alumnos; se trata de examinados muy capaces y acostumbrados a este tipo de pruebas.

preguntas, un recurso fácil es incluir como respuestas una de estas dos: *todas las anteriores* o *ninguna de las anteriores.*

La respuesta *todas las anteriores* no es recomendable porque puede construir una pista importante para desechar algunas alternativas. Si un alumno sólo sabe que una de las alternativas propuestas es verdadera, sólo tiene que *adivinar* entre esa respuesta y *todas las anteriores*.

Menos problema ofrece el uso de *ninguna de las anteriores* como respuesta. Los diversos estudios hechos sobre esta respuesta (Crehan, 1989; Knowles y Welch, 1992)[15] llegan a las mismas conclusiones:

1ª Las preguntas en las que se incluye *ninguna de las anteriores* como una respuesta más tienden a ser más difíciles; en esto concuerdan casi todos los estudios experimentales; en alguno de estos estudios (Tollefson, 1987) la pregunta es más difícil de hecho cuando esta respuesta es la respuesta *correcta.*

2ª Estas preguntas son también ligeramente menos discriminantes (diferencian menos entre los que más y menos saben).

3ª Por lo que respecta al test completo, estas preguntas hacen que la fiabilidad baje algo; en algún estudio en particular (Tollefson, 1987) aparece con claridad que la fiabilidad baja bastante cuando se comparan exámenes con las mismas preguntas y que sólo difieren en el uso frecuente de esta respuesta.

4ª Al menos un estudio (Dochy y otros, 2001)[16] muestra que en los ítems que presentan pequeños problemas con respuestas numéricas, esta respuesta *(ninguna de las anteriores)* atrae a los que no confían en la que hubiera sido su propia respuesta; cumple por lo tanto con su función.

Esta respuesta *(ninguna de las anteriores)* se puede aconsejar:

a) en lugar de *distractores* malos, a falta de otros mejores,
b) en no más de una cuarta o quinta parte de los ítems,
c) debe ser la respuesta correcta en una proporción similar (en la cuarta o quinta parte de los ítems que tienen esta opción),

[15] Crehan (1989) además de presentar un análisis propio recoge los resultados de 11 investigaciones sobre la respuesta *ninguna de las anteriores;* Knowles y Welch (1992) revisan 12 estudios y llegan a la conclusión de que esta respuesta no hace disminuir la calidad del test.
[16] Con una muestra de N = 169 universitarios en un examen de Ciencias.

d) en preguntas relativamente difíciles y en las que hay claramente una única respuesta correcta.

Rich y Johanson (1990) recomiendan que preguntas con esta alternativa, *ninguna de las anteriores*, aparezcan entre las primeras preguntas del test, y donde la repuesta correcta no se preste a confusión; de esta manera esta respuesta la verán los alumnos como más creíble.

3. Número de respuestas

El determinar cuál es el número de respuestas óptimo es importante. Intuitivamente podemos ver que a mayor número de respuestas va a ser más difícil adivinar la respuesta correcta; por otra parte un mayor número de respuestas supone un mayor esfuerzo por parte del que construye el test; no todas las alternativas falsas (o *distractores)* son siempre plausibles para el que no sabe; en definitiva un mayor número de respuestas no implica automáticamente una mayor calidad del test. Sobre este punto hay mucho investigado y aunque en estos temas nunca hay conclusiones definitivas para todas las situaciones, vamos a ver que en conjunto el número óptimo de respuestas es de tres; la correcta y dos alternativas incorrectas.

3.1. Recomendación y práctica habitual

La mayoría de los textos recomiendan cuatro o cinco alternativas (una correcta y tres o cuatro falsas o distractores). La razón que suele aducirse es que con un mayor número de alternativas disminuye la probabilidad de adivinar la respuesta correcta. El no recomendar más cuatro alternativas incorrectas se debe a la dificultad de redactar respuestas falsas y a la vez plausibles. Posiblemente cuatro es el número más frecuente, a pesar de los numerosos estudios (como iremos viendo) que favorecen la inclusión de sólo tres respuestas (la correcta y dos falsas).

A pesar de que son numerosos los estudios que confirman que, como criterio general, tres respuestas son suficientes (al menos preferible por el ahorro de tiempo que supone y sin desventajas claras), la mayoría de los textos siguen recomendando rutinariamente utilizar cuatro o cinco respuestas. Owen y Froman (1987) revisan 35 textos de evaluación, y 31 recomiendan utilizar más de tres respuestas sin citar ninguna investigación[17].

[17] La constante recomendación de utilizar más de tres respuestas (cuatro o cinco) a pesar de los muchos estudios que muestran que tres respuestas es suficiente (o mejor), les lleva a estos autores (Owen y Froman, 1987) a preguntarse *What's Wrong With Three-Option Multiple Choice Items?*. Estos autores, en su propia investigación utilizando tres y cinco alternativas, llegan a la conclusión de que la única diferencia importante está en el ahorro de tiempo que supone utilizar tres opciones.

3.2. Número de respuestas y discriminación en el conjunto de la muestra

Diversos estudios (sobre todo Lord, 1977a, 1977b; Levine y Drasgow, 1983; Owen y Froman, 1987, y otros que iremos mencionando) muestran que el número óptimo es de tres alternativas, una correcta y dos incorrectas. Es útil caer en la cuenta de que el número de respuestas discrimina de manera diferencial en los diversos segmentos de la muestra.

Los estudios de Lord, y de los otros autores mencionados, muestran que en general:

a) Dos alternativas (una correcta y otra incorrecta, o el clásico *verdadero-falso)* discriminan mejor *solamente* en la *parte alta* de la distribución: quedan más diferenciados los que saben más (que son quienes menos responden al azar) pero en el resto de la distribución quedan todos más indiferenciados.

b) Tres alternativas discriminan e informan mejor en el *centro* de la distribución (los mejores y los peores quedan menos diferenciados entre sí).

c) Cuatro o más alternativas dan mejores resultados en la *parte más baja* de la distribución, donde el adivinar es más frecuente; para los alumnos con menos ciencia las alternativas falsas pueden ser más plausibles y tienen más dónde escoger. Los que menos saben tienden más a adivinar, a mayor número de alternativas tienen más oportunidades de equivocarse y *quedan peor*, obviamente. Lo que sucede es que los que menos saben pueden quedar suficientemente diferenciados (quedar en el lugar que les corresponde) con menos alternativas.

3.3. Número de respuestas y fiabilidad de todo el test.

Entendemos fiabilidad en su sentido psicométrico: a mayor fiabilidad los sujetos quedan mejor diferenciados, mejor ordenados; en exámenes equivalentes se hubiera mantenido un orden semejante.

Resumimos brevemente las conclusiones de algunos estudios experimentales importantes.

1. Haladyna y Downing[18] (1985) revisan 56 estudios experimentales sobre la redacción de los ítems y añaden otro propio (Haladyna y Downing 1988, con N = 1111, estudiantes de medicina) llegando también a la conclusión de que en conjunto el número óptimo es el de tres alternativas: la fiabilidad tiende a subir al aumentar el número de alternativas (se discrimina

[18] Estos autores tienen numerosas investigaciones sobre las pruebas objetivas (en exámenes de Medicina) y por lo general hechas con muestras muy numerosas.

mejor en la parte más baja de la distribución), pero a partir de tres alternativas el aumento de la fiabilidad es mínimo y negligible.

2. Otros estudios no muestran una relación apreciable entre fiabilidad y número de respuestas. Trevisan, Sax y Michael (1994) concluyen que no hay diferencias en fiabilidad utilizando tres, cuatro y cinco respuestas, por lo que son preferibles tres opciones que además facilitan el aumentar el número de preguntas[19]. La no relación entre número de respuestas y características psicométricas (fiabilidad, validez) también la confirma el estudio de Rogers y Harley (1999): los tests con tres alternativas de respuesta son por lo menos equivalentes a los de cuatro respuestas.

En un estudio anterior (Trevisan y Sax, 1990) se analiza la relación entre número de alternativas y fiabilidad y validez de todo el test pero teniendo en cuenta la *capacidad intelectual* del alumno. Las conclusiones finales vienen a ser las mismas; en conjunto todos los resultados y para todos los grupos favorecen las *tres alternativas* como el formato *más eficaz.* A la misma conclusión llegan Bruno y Dirkzwager (1995), que además mencionan otros muchos estudios, analizando el problema desde otras perspectivas no puramente psicométricas: la información que aporta un test sube al aumentar el número de alternativas, pero a partir de tres el aumento es marginal (la calidad de las alternativas que no son igualmente atractivas; el tiempo y coste adicional no compensa aumentar el número de alternativas).

3. Al menos un estudio concreto (Cizek, Robinson y O'Day, 1998) muestra que al bajar de cinco a cuatro alternativas (eliminando la alternativa menos funcional) aumenta la discriminación de los ítems y la fiabilidad de todo el test[20]. Por otra parte la reducción en tiempo de examen puede permitir con facilidad ampliar el número de preguntas.

Ya vemos que no todos los estudios nos dan resultados consistentes, pero la conclusión obvia es que la investigación muestra que en general los tests de cuatro alternativas (o más) no son superiores a los de tres alternativas en términos de fiabilidad. Prácticamente todas las investigaciones mencionadas tienden a favorecer las tres alternativas.

3.4. Número de respuestas, *eficiencia* del test y calidad de los distractores

Por *eficiencia* se entiende aquí la razón tiempo/información obtenida. La máxima información en menor tiempo se obtiene con tres, o incluso cuatro, alternativas. Más

[19] En este estudio en las tres versiones de la prueba todos los ítems tienen tres opciones idénticas.
[20] En esta investigación 719 alumnos respondieron a 32 ítems en la versión de cinco respuestas y 726 respondieron la versión de cuatro respuestas

alternativas por ítem supone más tiempo de lectura y de contestar al test sin que compense la información adicional obtenida. La *eficiencia* suele ser mayor con tres alternativas, aunque, naturalmente, hay que procurar que todas las alternativas sean de calidad.

La calidad consiste fundamentalmente en que los dos distractores falsos sean *funcionales;* las dos características que proponen Haladyna y Downing (1988) para juzgar como funcionales a los distractores falsos son:

1º Que sean escogidos por más del 5% de la muestra;

2º Que tengan una correlación *(biserial-puntual)* negativa con el total; este análisis supone tratar cada alternativa como si fuera una pregunta, con respuesta 0 ó 1[21].

Esta correlación de cada ítem con el total *(menos el ítem)* es lo que se hace habitualmente con la alternativa correcta para determinar en qué grado discrimina cada ítem, *pero además* es útil hacerlo con todas las alternativas (sean tres o más) porque nos dice en qué medida el escoger una alternativa falsa está relacionado con estar bien o mal en el conjunto del test.

Cuando hay más de dos alternativas falsas, raramente *funcionan* más de dos, sobre todo en los niveles medios y superiores; todo esto se puede comprobar en cada ocasión mediante los análisis apropiados[22].

3.5. Número de respuestas y dificultad en redactar buenos distractores

Podemos añadir una razón más para no incluir muchas alternativas incorrectas: la dificultad en encontrar una tercera, cuarta o quinta alternativa que sea incorrecta y a la vez plausible, de manera que funcione eficazmente como distractor.

Esta dificultad la observan muchos autores como por ejemplo Heywood (1989) tratando sobre la evaluación en la universidad, y Bruno y Dirkwager (1995) que juzgan extremadamente difícil *(extremely difficult)* redactar ítems con cuatro o cinco respuestas igualmente atractivas; Burton (2001) advierte que a la dificultad de añadir una nueva alternativa que no sea obviamente incorrecta se añade el hecho de que pasar de cuatro a cinco respuestas tampoco afecta mucho a la fiabilidad de todo el test. Es además lo que suele observarse con mucha frecuencia cuando se analizan las respuestas; en muchas preguntas hay

[21] Cada alternativa de cada pregunta puntúa 1 si es escogida y 0 si no es escogida; lo que suponemos es que el escoger una alternativa incorrecta está relacionado con un total más bien bajo en todo el test.

[22] Rogers y Harley (1999:240) dan una serie de normas específicas y prácticas sobre qué alternativas se deben eliminar para pasar de cuatro alternativas a tres (a partir de los análisis hechos con la versión de cuatro alternativas).

distractores que nadie o casi nadie escoge, ni siquiera los que a juzgar por el mismo test que se analiza saben menos[23].

3.6. Número de respuestas y habilidad para responder correctamente sin conocer la respuesta correcta

Rogers y Harley (1999) incluyen una nueva variable en su investigación sobre el número de alternativas y que de alguna manera coincide con la dificultad anterior, sobre la dificultad de redactar buenas alternativas cuando las falsas son más de dos. El estudio de Roger y Harley se refiere la habilidad de algunos examinados para responder correctamente o adivinar con más facilidad la respuesta correcta sin conocimientos suficientes, simplemente eliminando las alternativas menos plausibles; ya hemos tratado también de este punto en otro contexto (sobre las *normas* para redactar buenas preguntas). La norma general para bajar de 4 (o más alternativas) a 3, es eliminar aquellas alternativas incorrectas escogidas por un menor número de alumnos (lo que supone haber hecho previamente un *análisis de ítems*).

Cuando los alumnos tienen solamente conocimiento parcial, esta habilidad (que no todos los examinados tienen en idéntico grado) contribuye a mejorar su puntuación total independientemente de su ciencia. Lo que estos autores confirman es que al reducir el número de alternativas (pasar de cuatro a tres) desaparece o disminuye el influjo de esta habilidad (o *astucia)* en la cual pueden diferir alumnos con idéntico conocimiento parcial; en definitiva al reducir el número de alternativas se eliminan las de peor calidad (a veces son absurdas) y desaparecen (o disminuyen) las pistas que facilitan responder correctamente sin conocimiento seguro.

Trataremos también sobre el influjo *legítimo* del conocimiento parcial en responder correctamente (lo veremos más adelante en el contexto de la *adivinación);* de lo que se trata aquí es de *eliminar pistas* (que básicamente son alternativas incorrectas de mala o dudosa calidad) que ayudan a unos alumnos más que otros al margen de que sepan más o menos.

3.7. Recomendaciones sobre el número de respuestas

Haladyna y Downing (1988) proponen en definitiva tres alternativas (una correcta y dos falsas) como número óptimo, y ya vemos que es una recomendación tan habitual como quizás poco seguida. Esta conclusión, la de utilizar tres respuestas aumentando el número de ítems,

[23] Cuando se analizan los ítems de una prueba objetiva, es normal calcular la correlación de cada ítem con el total, y diversos índices de dificultad y discriminación. Es de interés también comprobar cuántos de los alumnos del 25% con un total más alto y del 25% con un total más bajo escogen cada alternativa; este sencillo análisis pone de manifiesto de manera intuitiva (y sin necesidad de cálculos estadísticos que no todos entienden) el *atractivo* que pueden tener los diversos distractores.

viene avalada por el meta-análisis de Rodriguez (2005)[24]; no se alarga el tiempo necesario para responder, se obtiene una mayor información y no sufren las propiedades psicométricas de todo el test. Tampoco se *adivina más* cuando hay solamente tres opciones porque frecuentemente, cuando hay más de tres respuestas posibles, muchas de estas respuestas son poco plausibles.

Las razones para limitarse a tres respuesta, una verdadera y dos falsas, podemos resumirlas así:

1) Se ahorra tiempo en la preparación del test,
2) Se reduce la longitud del test,
3) Se reduce el tiempo de administración del test,
4) Se pueden mantener las características deseables en todo test (fiabilidad, información óptima sobre los examinados).

La razón más importante es la última; es una condición para poder aceptar las tres primeras razones, pues en ningún caso debe bajar la calidad del test.

Como conclusiones razonables de lo mucho investigado sobre el número de alternativas podemos añadir:

a) Parece claro que al preparar un test objetivo la *norma* debería ser utilizar tres respuestas y aumentar, en cambio, el número de ítems para *cubrir* más contenido, aumentando también la fiabilidad (que aumenta al aumentar el número de ítems): son preferibles cuatro preguntas con tres alternativas que tres preguntas con cuatro alternativas.

b) Si en una primera versión de un test con cuatro o cinco alternativas, ya preparado y utilizado, se observan alternativas *no funcionales* (a juicio de expertos, o apenas escogidas por los alumnos, etc.), es preferible disminuir el número de opciones en las versiones subsiguientes. Los *análisis de ítems* (de los que no estamos tratando aquí) son muy útiles para ir mejorando la calidad de estas pruebas.

[24] Para este meta-análisis (Rodriguez, 2005, que presenta una *historia* completa de estos análisis) el autor seleccionó, de los 48 estudios empíricos localizados, los 27 que cumplían estas dos condiciones, evaluar el efecto diferencial al variar el número de opciones de respuesta y aportar toda la información relevante (número de sujetos, número de ítems en cada formato, fiabilidad, etc.). En todos los casos se trata de estudios experimentales con asignación aleatoria de los participantes a las diversas condiciones (con tres o más respuestas). La media del número de ítems en las pruebas revisadas es de 43, y el número total de sujetos es de 12.591. La fiabilidad se reduce en los casos en que se baja de 5 ó 4 respuestas a 2, pero no cuando se pasa de 4 a 3 respuestas.

c) Podemos añadir además que es importante analizar las pruebas objetivas que de hecho se utilizan para irlas mejorando; estos análisis pueden encontrarse ya programados y son de fácil comprensión.

Esta orientación (más preguntas y menos alternativas) puede ser especialmente conveniente en exámenes que interesa que sean más bien largos.

4. Preguntas del tipo Verdadero-Falso

Estas preguntas son tan frecuentes como discutidas. Es normal juzgar este tipo de preguntas como de peor calidad que las que tienen más posibles respuestas. Las limitaciones que suelen señalarse a estas preguntas son a) que el adivinar pesa mucho, b) que comprueban sobre todo conocimientos de memoria y c) que con frecuencia son ambiguas (porque si se formula la afirmación con mucha claridad y matiz, se puede convertir en claramente falsa o verdadera).

a) El que el adivinar pese mucho en la puntuación total dependerá del número de preguntas y de dónde se ponga el número mínimo de respuestas correctas para el apto. En cualquier caso la probabilidad de responder correctamente a una pregunta respondiendo al azar es del 50%.

Para disuadir un adivinar ciego se han propuesto con éxito (Gardner-Medwin, 1995) otros modos de corrección de estas preguntas, incluyendo para cada pregunta otra pregunta en la que el alumno debe indicar su grado de seguridad en la respuesta. La puntuación que recibe el alumno no depende solamente de que su respuesta sea correcta o incorrecta, sino de su grado de seguridad en que la respuesta sea correcta (si la respuesta es incorrecta: poca seguridad = 0, cierta seguridad = -2 y mucha seguridad = -6; si la respuesta es correcta: los valores son 1, 2 y 3 respectivamente). Este punto lo tratamos con más extensión en el apartado 6.3, sobre las alternativas al uso de la formula de corrección por adivinación.

b) El que estas preguntas comprueben preferentemente conocimientos de pura memoria también es discutible; depende de cómo se formulen, aunque es verdad que son más fáciles de componer para comprobar conocimientos de memoria. Para autores de indudable prestigio y experiencia en este campo, como Ebel (1977) que comenta ampliamente este tipo de preguntas y da normas sobre su redacción, estas preguntas pueden ser útiles, discriminantes y pueden además estimular la capacidad de pensar del alumno *(thought provoking).* En opinión de este autor cuando estas preguntas no son buenas preguntas se debe a que no están bien pensadas y redactadas más que al mismo formato de Verdadero-Falso.

c) El que con frecuencia sean ambiguas quiere decir que no es tan fácil como puede parecer el redactarlas bien. Es verdad que muchas de estas preguntas (afirmaciones) son verdaderas o falsas según se den o no se den determinadas condiciones que no siempre se especifican, por lo que pueden resultar ambiguas para muchos alumnos. En definitiva es un problema de cuidado en la redacción de estas preguntas.

Sobre estas preguntas se ha investigado mucho acerca de su facilidad y dificultad y de su poder discriminatorio; también hay estudios que las comparan según diversos criterios con las preguntas de elección múltiple y se han propuesto alternativas para minimizar sus limitaciones. Sobre estos puntos tratamos en los apartados siguientes.

4.1. La dificultad de las preguntas con respuesta Verdadero-Falso

La dificultad de estas preguntas depende de dos factores:

1º que la *respuesta correcta* sea *verdadero* o *falso*,
2º que el ítem esté formulado *positivamente* o *negativamente*.

Varios estudios experimentales examinan la dificultad relativa de los ítems según estos dos factores, y es frecuente encontrar que:

formulación del ítem	***más fácil** cuando la respuesta correcta es*	***más difícil** cuando la respuesta correcta es*
positiva	Verdadero	Falso
negativa	Falso	Verdadero

Naturalmente no hay que esperar unanimidad en estos estudios (Wason, 1961; Peterson y Peterson, 1976). Lo más claro parece ser que los ítems con formulación negativa son de hecho más difíciles cuando la respuesta correcta es *verdadero*.

Sobre la formulación negativa, y por lo que respecta a cuestionarios sociológicos o sobre actitudes, los sujetos cambian con frecuencia su primera respuesta cuando la formulación del ítem incluye *no* o *nunca* (Edvarson, 1980); estas confusiones son más frecuentes en niños (Marsh, 1986).

4.2. El poder discriminatorio de las preguntas Verdadero-Falso

Los ítems cuya respuesta correcta es *Falso* suelen ser *más discriminantes* y tienen por lo tanto una *fiabilidad mayor* (considerados como un sub-test) que los que tienen *Verdadero* como respuesta correcta.

La razón está en que cuando se responde con duda o ignorancia intentando adivinar la respuesta correcta es más frecuente elegir *Verdadero* como respuesta. Los que saben más y los que saben menos quedan más diferenciados en las preguntas con respuesta correcta *Falso* y más igualados cuando la respuesta correcta es *Verdadero*. La tendencia a responder *Falso* de manera preferente es menos usual (Cronbach, 1942; Larkins y Swint, 1976; Grosse y Wright, 1985). Textos importantes de medición educacional recomiendan por este motivo incluir más preguntas (hasta un 67%) cuya respuesta correcta sea *Falso* (Ebel, 1977: 231).

4.3. Los tests del tipo Verdadero-Falso comparados con los de elección múltiple:

1º Tienden a ser:

a) Más fáciles (se acierta más adivinando),
b) Menos discriminantes
c) La fiabilidad de todo el test es menor manteniendo constante el número de ítems.

Esta menor fiabilidad de los tests con preguntas de dos alternativas con respecto a los que tienen tres respuestas o más (manteniendo constante el número de preguntas) está muy comprobada experimentalmente (Frisbie, 1973; Ebel, 1975; Lord, 1977a; Straton y Catts, 1980; Grosse y Wright, 1985).

2º Para conseguir una fiabilidad suficiente y reducir el influjo de la adivinación hacen falta más ítems (5 ítems *Verdadero-Falso* por cada 3 de elección múltiple para obtener una fiabilidad comparable). Con sólo dos alternativas (*Verdadero-Falso* u otro par de alternativas) se han sugerido hasta 150 ítems para conseguir una fiabilidad lo suficientemente adecuada como para tomar decisiones sobre los examinados (Downing, 1992).

3º Se responden en un tiempo menor; típicamente se responden tres ítems del tipo *Verdadero-Falso* en el mismo tiempo en que se responden dos de elección múltiple (Frisbie, 1973; Downing, 1992).

4º Los tests con preguntas de sólo dos respuestas suelen tener con otros criterios correlaciones semejantes a las que tienen tests con preguntas de más respuestas (citas de Downing, 1992).

4.4. Conversión de preguntas de elección múltiple en preguntas Verdadero-Falso

Se puede pensar en convertir las preguntas de *elección múltiple* (varias respuestas, una correcta) en preguntas del tipo *Verdadero-Falso* (cada alternativa se convierte en una pregunta) al menos en dos circunstancias:

1º Cuando las preguntas de *elección múltiple* ***equivalen de hecho*** a una serie de afirmaciones independientes a las que se podría responder directamente *Verdadero* o *Falso*. En este caso puede ser preferible convertir las alternativas en otras tantas preguntas del tipo *Verdadero-Falso* (Ebel, 1978; Frisbie y Sweeney, 1982):

a) Suelen tener mayor fiabilidad porque en este caso aumenta notablemente el número de ítems (si la fiabilidad es alta se puede pensar que el *adivinar* no ha pesado mucho, Ebel, 1978); de todas maneras el problema que puede suponer el adivinar lo tratamos más adelante;

b) Se responden en menos tiempo;

c) Son más fáciles de construir.

En cualquier caso las preguntas de *elección múltiple* cuyas respuestas son una serie de afirmaciones independientes no son en principio buenas preguntas de elección múltiple..

2º Cuando las preguntas de *elección múltiple tienen varias respuestas correctas* pueden convertirse con ventaja en *bloques* de preguntas del tipo *Verdadero-Falso.* Se utiliza así el *conocimiento parcial* del alumno que puede conocer unas respuestas y desconocer otras. El tener en cuenta este conocimiento parcial aumenta la fiabilidad y validez de todo el test, como veremos más adelante.

El problema de las dos respuestas, Verdadero-Falso u otras, está en la mayor probabilidad de responder sin saber y acertar con la respuesta correcta (con dos respuestas y respondiendo al azar, se puede responder correctamente el 50% de las preguntas). Más adelante tratamos cómo se pueden corregir y puntuar estas preguntas (bloques de afirmaciones *Verdadero-Falso)* para evitar el influjo del adivinar; se puede utilizar una clave de corrección parecida a ésta: todas las respuestas correctas = 1, más de la mitad = .50 y menos de la mitad = 0; en vez de asignar una puntuación de .25 a cada respuesta correcta en el caso de que hubiera cuatro alternativas (cada *bloque* de cuatro o cinco alternativas constituye una pregunta de valor máximo = 1).

4. 5. Conversión de los ítems Verdadero-Falso en preguntas de elección múltiple

Cuando lo permite el tipo de pregunta, se puede evitar la dicotomía *Verdadero-Falso* al menos de dos maneras:

a) Proponiendo *varias respuestas* desde el *claramente verdadero* al *claramente falso*, sobre todo a propósito de ítems (proposiciones) que tienen un cierto grado de complejidad, así las respuestas podrían ser (Heywood, 1977):

A ☐	Claramente verdadero	C ☐	Sin información suficiente para justificar su veracidad
B ☐	Posiblemente verdadero	D ☐	Claramente falso

Estas respuestas pueden ser otras que se consideren más idóneas, menos ambiguas, etc. Las afirmaciones propuestas como preguntas pueden a veces ser verdaderas o falsas según se den distintas circunstancias, según los datos presentados para apoyar su supuesta veracidad, etc. y se pueden transformar en preguntas con más de dos alternativas.

b) Otra manera de formular estos ítems de manera que inviten a un *adivinar inteligente*, y evitar dos únicas respuestas, consiste en presentar dos afirmaciones (con tal de que *tenga sentido* presentarlas juntas) con estas cuatro respuestas (Eakin, 1977):

A ☐	Las dos verdaderas	C ☐	Las dos falsas
B ☐	Sólo la 1º verdadera	D ☐	Sólo la 2º verdadera

Se puede pensar en otros tipos de preguntas a partir de dos afirmaciones, como es identificar relaciones entre ambas (una puede ser una consecuencia de la otra, o una condición, etc.).

4.6. Preguntas de *respuesta alternativa*

Son preguntas con sólo dos respuestas pero distintas del *Verdadero-Falso*. Aunque coincidan con las preguntas del tipo *Verdadero-Falso* en que sólo hay dos respuestas, sus características son distintas, por eso es útil distinguirlas.

Las respuestas *Verdadero-Falso* se pueden sustituir por otras dos alternativas formuladas de otra manera; por ejemplo (Ebel, 1982):

Los eclipses de sol solamente pueden ocurrir en luna: ☐ nueva ☐ llena

en vez de

Los eclipses de sol solamente pueden ocurrir en luna nueva: ☐ verdadero
☐ falso

Estas preguntas pueden ser una buena alternativa al *Verdadero-Falso* (Ebel, 1982, 1983):

a) Se puede evitar con más facilidad la ambigüedad inherente a muchas preguntas del tipo *Verdadero-Falso* porque permiten comparar las dos respuestas sin requerir un juicio absoluto sobre la falsedad o veracidad de una proposición;

b) También posiblemente se redactan con mayor facilidad (no es siempre fácil redactar proposiciones *absolutamente* verdaderas o falsas);

c) Permiten con facilidad comprobar conocimientos que no son de pura memoria.

Estas preguntas además suelen ser *más discriminantes* y consecuentemente los tests hechos con estas preguntas tienen una mayor fiabilidad si los comparamos con los del tipo *Verdadero-Falso* (Downing, 1992). Estas preguntas se asemejan más a las preguntas de elección múltiple; en realidad son preguntas de elección múltiple con sólo dos respuestas, e invitan menos a un adivinar ciego. De hecho en muchos tests con preguntas de elección múltiple sólo hay uno o dos distractores funcionales (atractivos para el que no sabe). Naturalmente el problema de acertar adivinando es el mismo que en las preguntas *Verdadero-Falso.*

4.7. Preguntas Verdadero-Falso indicando el nivel de seguridad de la respuesta

Este punto lo tratamos con más extensión en el contexto de la adivinación pero es conveniente mencionarlo aquí pues tiene especial aplicación en las preguntas del tipo Verdadero-Falso que pueden quedar revalorizadas con estos sistemas.

Con este formato los alumnos, además de responder Verdadero o Falso, tienen que indicar su nivel de confianza (1 = bajo, 2 = moderado, 3 = alto) en haber respondido correctamente, de manera que responder incorrectamente y con cierta seguridad queda penalizado. Este sistema se ha utilizado con éxito al menos en Medicina.

El valor de la respuesta propuesto para las preguntas Verdadero-Falso es (Gardner-Medwin, 1995):

Grado de confianza en la respuesta correcta:

	Bajo (1)	Medio (2)	Alto (3)	Sin respuesta
Respuesta correcta	1	2	3	0
Respuesta incorrecta	0	-2	-6	0

Naturalmente puede haber otras maneras de valorar estas respuestas, pero el penalizar las respuestas incorrectas dadas con un cierto nivel de seguridad tiene un efecto claramente disuasorio para lo que es el mayor problema de estas preguntas (el *adivinar* la respuesta correcta), y a la vez *se premia* el nivel de seguridad si la respuesta es correcta y este nivel de seguridad se expresa con sinceridad. Más comentarios y explicaciones sobre este tipo de respuestas pueden verse en el apartado 6.3, en el contexto de las alternativas a la fórmula de corrección por adivinación.

5. Preguntas con varias respuestas correctas

La forma habitual de presentar las preguntas objetivas es con una sola respuesta correcta, pero a veces la única respuesta correcta es una combinación de respuestas correctas, o se responde de manera independiente a varias preguntas del tipo *Verdadero-Falso* que componen una única pregunta. También se utilizan en un sentido más literal: el alumno debe escoger la combinación correcta de posibles respuestas. En los apartados siguientes tratamos sobre estas modalidades, sus ventajas relativas y algunos modos específicas de corrección de estas preguntas.

Estos formatos más complejos que el tradicional *(una única respuesta correcta)* han merecido en los últimos años una especial atención para responder a la crítica generalizada (y fundada) de que las pruebas objetivas convencionales no comprueban (y por lo tanto no estimulan) objetivos que implican por parte del alumno el uso de su capacidad de pensar, relacionar, etc. *(higher-order thinking).*

Esta limitación de las preguntas objetivas más frecuentes es lo que, por ejemplo, comprueban entre otros, Bateman y Kato (1993, que aducen además otros muchos estudios)[25] refiriéndose a niveles universitarios: los tests habituales miden preferentemente información,

[25]Estas investigadoras, del *Centre for University Teaching and Learning* de McGill University, Montreal, estudian 27 cursos correspondientes a otros tantos profesores de carreras de Ciencias Sociales (en este grupo se incluyen *Bussiness Administration*, Económicas, Ciencias Políticas, Psicología, Sociología…). Analizan un total de 2275 ítems o preguntas de evaluación. En conjunto conocimiento y comprensión (los niveles más bajos) comprenden el 80% de los ítems, preguntas o ejercicios diversos, incluyendo los exámenes finales; un resultado que corrobora el de otros muchos estudios. Si se tienen en cuenta solamente las preguntas objetivas, el porcentaje de los ítems que comprueban sólo conocimientos y comprensión son el 93% (y el 66% de las preguntas no objetivas).

comprensión, etc. pero sin llegar a objetivos de más calibre (capacidad de síntesis, de evaluación, etc.).

Una razón es que estas preguntas no se prestan para comprobar este tipo de objetivos más complejos (ciertamente no los que se refieren a la capacidad de organización y otros que requieren pruebas abiertas), pero otra razón importante es que la formulación de preguntas objetivas que midan objetivos más complejos son más difíciles de construir. En esta categoría de preguntas objetivas más complejas entran las que tratamos ahora con varias respuestas correctas, aunque los ejemplos que ponemos aquí como ejemplo de los diversos formatos son muy sencillos (por razones de simplicidad en la presentación).

Estos formatos con varias respuestas correctas, que se prestan a exigir al alumno que compare, detecte relaciones, etc., no son de ahora; ya fueron estudiados por Cronbach (1941)[26] hace años; lo que sí es de ahora es el nuevo énfasis en este tipo de preguntas, buscando en el alumno un estudio más inteligente y una comprobación de algo más que memoria y simple comprensión (Glasnapp y Poggio, 1994).

5.1. Diversas maneras de presentar las preguntas con varias respuestas correctas

Los ítems con varias respuestas correctas admiten distintas formas de presentación, de modos de responder (y consecuentemente de normas que se dan a los alumnos) y de corrección.

Todos estos diversos formatos, más complejos que el habitual de una única respuesta correcta, han sido ampliamente investigados en EE. UU., sobre todo en los exámenes finales de Medicina (en algunos estudios que mencionamos aquí se analizan miles de alumnos y cientos de preguntas). Sus ventajas, inconvenientes y peculiaridades están muy investigadas, al menos en el caso de alumnos universitarios, de nivel por lo general alto y habituados a responder a tests. Las conclusiones de estos estudios dan una buena idea sobre cómo suelen *funcionar* los diversos formatos.

En las figuras 1, 2 y 3 tenemos tres ejemplos sencillos de los tipos de formato que suelen adoptar estas preguntas (que suelen tener cuatro o cinco respuestas).

[26] Las preguntas objetivas con más de una respuesta correcta se vienen investigando desde los años 40 y 50, sobre todo comparando los dos formatos que veremos a continuación, el de *varias respuestas correctas* y el *múltiple verdadero-falso;* pueden verse citas en Pomplum y Omar (1997).

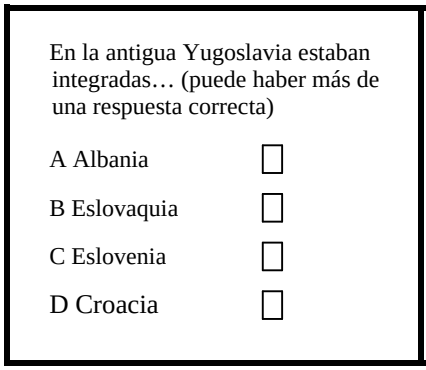

Fig. 1
Varias Respuestas Correctas

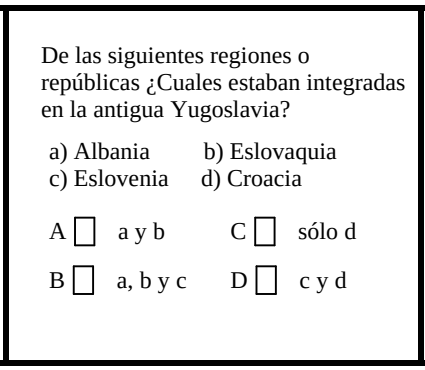

Fig. 2.
Elección Combinada

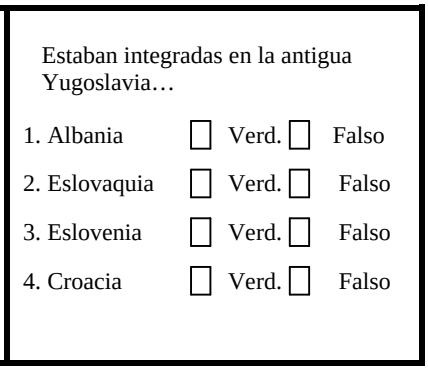

Fig. 3.
Múltiple Verdadero-Falso

Estos formatos no sólo se diferencian entre sí por el tipo de respuesta. La misma pregunta o afirmación inicial puede responder a dos tipos distintos:

a) Puede ser una verdadera pregunta (o su equivalente, una afirmación incompleta) con todas las respuestas pertenecientes al mismo ámbito y presentadas como posibles respuestas a la pregunta inicial del bloque. El alumno debe comparar unas respuestas (o subpreguntas) con otras para responder correctamente.

b) También se pueden formular de manera que la pregunta sea un simple encabezamiento para el resto de las posibles preguntas, que pueden ser muy independientes entre sí aunque dentro del mismo tema general. La forma más sencilla sería preguntar *¿Cuáles de las siguientes afirmaciones son verdaderas (o falsas)?*

No siempre es fácil la distinción entre los dos modos de preguntar, pero en principio parece preferible que las respuestas o subpreguntas pertenezcan al mismo ámbito y centren bien la atención del alumno.

5.2. Preguntas con *varias respuestas correctas* (fig. 1)

Al alumno se le instruye que puede haber más de una respuesta correcta (esta advertencia hay que hacerla siempre). En principio se considera que la pregunta está bien respondida si elige *todas y solas* las respuestas correctas. Como vamos a ver, estas preguntas son en general más difíciles que las preguntas de *elección combinada* (fig. 2).

Si en las preguntas de *múltiple Verdadero-Falso* (fig. 3) cada bloque se considera como una única pregunta (todo bien = 1, algo mal = 0) estamos en el mismo caso, aunque quizás el formato *Múltiple Verdadero-Falso* facilite la respuesta al alumno; de hecho las mismas preguntas con varias respuestas correctas presentadas con el formato de *múltiple Verdadero-*

Falso resultan algo más fáciles para los alumnos (y las de *elección combinada* las más fáciles, Frisbie, 1992).

Como advierten Glasnapp y Poggio (1994), este formato lleva un mensaje implícito al alumno que también es un *mensaje educativo:* no hay siempre una única respuesta correcta, sobre todo cuando se trata de verificar comprensión de un pasaje del que se pueden inferir legítimamente diversas inferencias y significados. Este mensaje queda diluido cuando estas preguntas, con varias respuestas correctas, se presentan con el formato de *Múltiple Verdadero-Falso* que examinaremos enseguida.[27]

Pomplun y Omar (1997) tienen una amplia investigación con este tipo de preguntas (unos 30.000 alumnos en exámenes de matemáticas y comprensión lectora) y concluyen que:

a) en términos de fiabilidad y validez son aceptables,
b) se controla mejor la adivinación que en el método alternativo de *múltiple verdadero-falso,*
c) son coherentes con buenos objetivos educacionales,
d) se prestan con facilidad a corrección mecánica o electrónica.

Sobre todo con niños, y más si están acostumbrados a *una única respuesta correcta,* las normas de respuesta tienen que ser muy claras y conviene que vean antes algún ejemplo.

5.3. Preguntas de *elección combinada* (fig. 2)

En las respuestas se presentan *combinaciones de posibles respuestas* previamente indicadas. En definitiva hay una sola respuesta correcta, es decir, una sola combinación correcta.[28]

Sobre estas preguntas se pueden hacer estos comentarios:

---

[27] Estos autores (Glassnapp y Poggio, 1994) analizan, en alumnos de secundaria, aproximadamente 15.000 respuestas a ítems de este formato. Se trata de preguntas de comprensión de textos con cinco posibles respuestas, y cada pregunta puntúa de 0 a 5, según el número de *decisiones correctas* del alumno (las decisiones correctas son escoger las respuestas verdaderas y *no* escoger las falsas). Estas preguntas funcionan peor (en términos de fiabilidad, de no entender las normas de respuesta, etc.) con los alumnos de edad más baja. Cuando de cada alumno se obtienen dos totales, la suma de elecciones correctas cuando la respuesta es *verdadero* y la suma de elecciones correctas cuando la respuesta es *falso* (la elección correcta aquí es no responder), la fiabilidad está en torno a .80 con las elecciones correctas cuando la respuesta es *verdadero*, y la fiabilidad está entre .60 y .70 con las elecciones correctas cuando la respuesta es *falso*. También hay correlaciones notablemente mayores entre elecciones correctas de respuestas verdaderas y conocimientos previos, y dos medidas de actitud hacia el estudio. La implicación que comentan los autores es que el tener en cuenta solamente las respuestas correctas a afirmaciones verdaderas (más válidas y fiables) puede ser una alternativa a las preguntas convencionales con una única respuesta correcta, aunque éste es un punto no suficientemente investigado.

[28] En la literatura sobre estos temas este tipo de ítems (las alternativas son combinaciones de respuestas) se denominan a veces *K-Type items*, y los de una simple respuesta correcta *A-Type items*, aunque no hay consistencia en el uso de esta terminología

1º Tienen una ventaja inicial: en realidad sólo hay una *única respuesta correcta* (una única combinación correcta de respuestas), y un número de alternativas por ítem que puede mantenerse constante a lo largo del mismo test, tanto si hay una sola respuesta correcta como si hay varias, y así no se rompe el esquema general. Se facilita además la corrección con lectura óptica o por otros medios, lo mismo que otros análisis estadísticos en los que se supone que hay una única respuesta correcta.

2º El *inconveniente mayor* de estas preguntas es que en las respuestas fácilmente hay *pistas* sobre cuál es la respuesta correcta. Si en el ejemplo de la fig. 2 un alumno sabe que Albania no formó parte de la antigua Yugoslavia, puede eliminar las alternativas A y B y adivinar solamente entre C y D.

3º Si comparamos el formato de *elección combinada* (fig. 2) con el de *varias respuestas correctas* (fig. 1, en el que se considera que una pregunta está bien respondida cuando se escogen todas y solas las alternativas correctas) y con el formato de *múltiple Verdadero-Falso* (fig. 3), estas preguntas de *elección combinada:*

a) son más fáciles,
b) el test en su conjunto tiene menor validez (correlaciones menores con otros criterios);
c) y también tiene menor fiabilidad.

Estas conclusiones aparecen en numerosos estudios (Haladyna y Downing, 1985; Albanese y Sabers, 1988; Albanese, 1993). Estos datos confirman que, en efecto, con este formato es casi inevitable dar pistas para escoger la respuesta correcta, al menos cuando se comparan con los otros formatos en los que hay también varias respuestas correctas.

4º Comparados con el formato más habitual *(una única y simple respuesta correcta),* estas preguntas de *elección combinada:*

a) tienden a ser *algo más difíciles* (Albanese, 1993; Subhiyah y Downing, 1993);

b) *requieren una mayor atención y tiempo* por parte de los alumnos, por lo que su eficiencia (relación tiempo requerido/información aportada) es menor; esto es quizás lo más importante como advierten varios investigadores sobre estos formatos (Haladyna y Downing,1985; Subhiyah y Downing, 1993).

En uno de estos estudios (Subhiyah y Downing 1993) se comparan 308 ítems, 154 del tipo convencional *(una única respuesta correcta)* y 154 de *elección combinada,* cubriendo los mismos contenidos y respondidos por un total de 7083 examinados. Los resultados no

muestran diferencias psicométricas importantes entre ambos tipos de formato, pero sigue siendo verdad que los ítems de *elección combinada* resultan ligeramente más difíciles y requieren más tiempo para ser respondidos cuando se comparan con el formato de una *única respuesta correcta*.

A pesar de los resultados más bien desfavorables de este tipo de ítems cuando se comparan con los de una *única respuesta correcta*, algunos constructores de tests los mantienen porque los juzgan especialmente adecuados para una determinada disciplina (medicina en este caso). Por otra parte los alumnos suelen preferir el formato más convencional (varias respuestas independientes, no una combinación de posibilidades, de las que sólo una es correcta).

Resumiendo podemos concluir que el formato de *elección combinada* o de una *combinación de respuestas correctas:*

a) Si lo comparamos con los otros de *varias respuestas correctas* (fig, 1, 2 y 3), es el más débil de los tres formatos; fácilmente se dan pistas sobre la respuesta correcta, son preguntas que tienden a ser más fáciles y menos discriminantes; además lleva más tiempo responder a estas preguntas que a las convencionales con una única respuesta correcta (no una combinación de respuestas).

b) Si lo comparamos con el formato de una *única respuesta correcta* (y esta suele ser la alternativa real para mantener en definitiva una sola respuesta correcta), la elección combinada suele ser *algo más difícil* para el alumno, lleva *más tiempo en responder*, pero no hay diferencias psicométricas importantes.

El elegir un tipo u otro de formato (*una simple* respuesta correcta o *una combinación* de respuestas correctas) dependerá habitualmente de lo que pida el contenido de la pregunta. Sí parece claro que no debe haber muchas preguntas de este tipo en el mismo test.

Una ventaja de estas preguntas es que son fáciles de componer y resulta un método sencillo de aumentar el número de distractores, manteniendo el *formato de una única respuesta correcta* con sus facilidades de corrección. A veces se redacta una pregunta y lo que con facilidad le viene a la mente del profesor son varias respuestas correctas, no varias incorrectas. Estas preguntas pueden ser por lo tanto un buen recurso para redactar respuestas incorrectas.

5.4. Preguntas de múltiple Verdadero-Falso (fig. 3)

Son preguntas con varias respuestas correctas pero, tal como se presentan al alumno, cada respuesta es una pregunta que se responde de manera independiente.

Estas preguntas *(múltiple Verdadero-Falso)* se pueden corregir de varias maneras:

a) Considerando cada *bloque* o *cluster* (como el de la fig. 3) como una sola pregunta. En este caso la pregunta inicial es una verdadera pregunta (o frase incompleta) que centra la atención del alumno sobre algo específico, y lo que tenemos es otra manera de presentar las preguntas con *varias respuestas correctas.* La pregunta se considera bien respondida si se responde correctamente, *Verdadero* o *Falso,* a cada una de las alternativas propuestas.

b) Se puede considerar cada subpregunta como una pregunta independiente, y en este caso cada *Verdadero-Falso* suele incluso llevar la numeración que le corresponde, como cualquier otro ítem del test. Estamos en la situación general de los ítems con dos respuestas, *Verdadero* o *Falso,* aunque presentados de otra manera (un encabezamiento cada cuatro o cinco ítems *Verdadero-Falso),* y con los mismos problemas (una probabilidad del 50% de adivinar la respuesta correcta).

c) Se puede considerar cada bloque o *cluster* como una única pregunta (valor máximo de la pregunta = 1) pero corrigiendo y puntuando de manera que se tenga en cuenta tanto el problema de la *adivinación* (la posibilidad de acertar respondiendo al azar), y como el *conocimiento parcial* del alumno, que puede conocer unas respuestas y no otras. Esto supone utilizar claves de corrección distintas según el número de aciertos dentro de cada ítem o bloque, como explicaremos enseguida.

Aunque los resultados de los diversos estudios hechos sobre este tipo de preguntas no son siempre coherentes y tampoco responden a todas las preguntas que uno se podría hacer, la investigación disponible (Frisbie, 1992 y Albanese, 1992 comentan numerosos estudios) nos indica algunas conclusiones que se pueden tener en cuenta.

1º La fiabilidad de los tests *múltiple Verdadero-Falso* en los que puntúa el *bloque entero* (1 ó 0) y no cada respuesta de manera independiente, tienden a tener una fiabilidad semejante o incluso más alta que los tests de elección múltiple con *una sola respuesta correcta*. Una fiabilidad más alta quiere decir en definitiva que se diferencian más claramente los que tienen totales más altos y más bajos; los alumnos quedan clasificados con más nitidez.

Downing, Grosso y Norcini (1994) comparan este formato *múltiple verdadero-falso* con el convencional de *una única respuesta correcta;* la muestra supera los 21000 alumnos

(exámenes para obtener la licencia en medicina), y los ítems analizados superan los 1000 de cada tipo. En todas las submuestras (los autores presentan 8 análisis) los exámenes del tipo *múltiple verdadero-falso* tienen una fiabilidad mayor que cuando hay una sola respuesta correcta. Sin embargo los coeficientes de validez (en este caso correlaciones con exámenes prácticos) favorecen ligeramente en algunos casos a las preguntas con una sola respuesta correcta. Otro estudio semejante (Baranowski y otros, 1994, con N = 10.867) muestra la misma tendencia; la fiabilidad es mayor con los ítems *múltiple verdadero-falso* (los candidatos quedan *mejor clasificados* en todas las habilidades cognitivas que se analizan por separado) pero las correlaciones con exámenes prácticos son algo mayores cuando se emplea el formato de una única respuesta correcta.

Aunque estos análisis nunca pueden considerarse conclusivos, estos investigadores (de amplia experiencia) tienden a preferir el formato tradicional de *una única respuesta correcta,* al menos en *estas situaciones* en las que se pretende de alguna manera *predecir* competencia profesional y medir habilidades cognitivas de cierto calibre.

2º La fiabilidad de los tests *múltiple Verdadero-Falso* en los que puntúa cada respuesta *Verdadero-Falso* como un ítem independiente también tienen una fiabilidad semejante o algo superior a los ítems de *una sola respuesta correcta,* pero con tal de que haya un número mayor de ítems *Verdadero-Falso* (una diferencia en torno de 3 a 1).

3º Cuando hay varias respuestas correctas, el formato de *múltiple Verdadero-Falso* resulta más fácil que el de *varias respuestas correctas* y más difícil que el de *elección combinada* (que además de ser el más fácil es el que se lleva más tiempo en responder).

4º La fiabilidad será en general distinta según se trate *cada alternativa* Verdadero-Falso como un ítem independiente (y aumenta el número de ítems y también por lo general la fiabilidad) o si se considera que cada *bloque* forma un único ítem con puntuación de 0 ó 1 (y disminuye el número de ítems).

La fiabilidad en estos casos no depende simplemente del número de ítems; la fiabilidad es menor si se trata el bloque como una pregunta cuando la varianza *dentro* de los bloques es mayor que la varianza *entre* los bloques (de sub-preguntas; en Albanese, 1993, puede verse discutido este punto).

Lo que sí es más claro es que si cada bloque forma una pregunta independiente, la fiabilidad es mayor si se tiene en cuenta el conocimiento parcial del alumno y se utilizan las claves de corrección que exponemos a continuación.

5º Una dificultad práctica de este formato la advierten varios autores (como Pomplun y Omar, 1997). Cuando se corrigen por medios mecánicos o de lectura óptica, la hoja de respuestas requiere muchos más espacios para responder (cada alternativa funciona como una pregunta a efectos de corrección), y si estas preguntas van mezcladas con las convencionales de una única respuesta correcta, se complica más la corrección. En cambio en las preguntas de varias respuestas correctas es válida la hoja de respuestas convencional.

5.5. Las preguntas *múltiple Verdadero-Falso,* el *conocimiento parcial* del alumno y el problema de la *adivinación*: alternativas a la corrección de estas preguntas

Cuando se trata de utilizar preguntas con *varias respuestas correctas*, de todo lo que vamos exponiendo se deduce que *parecen* ser preferibles las preguntas de *múltiple Verdadero-Falso* (fig. 3) aunque no es una conclusión clara ni de todos los autores. Si las comparamos con las preguntas con *varias respuestas correctas* (fig. 1) e incluso con las habituales con una única respuesta correcta, estas preguntas tienen una ventaja clara y un inconveniente también obvio.

La *ventaja* está en que permiten apreciar y tener en cuenta el *conocimiento parcial* del alumno, que puede saber unas cosas y no otras. Además son bastantes los estudios (citados por Albanese y Sabers, 1988) que muestran que el tener en cuenta el *conocimiento parcial* del alumno lleva a un aumento de la fiabilidad y validez de todo el test. Claro está que esta ventaja sólo se da si precisamente se tiene en cuenta este *conocimiento parcial* en el modo de corregir estas preguntas, como aclaremos enseguida.

El *inconveniente* está en la probabilidad mayor de *adivinar* la respuesta correcta en algunas de las *sub-preguntas* (sobre todo cuando la respuesta correcta es *verdadero);* esta probabilidad es mucho menor en el formato de *varias respuestas correctas* (fig. 1).

El formato de *múltiple Verdadero-Falso* se puede corregir de varias maneras de modo que por una parte se aproveche el *conocimiento parcial* del alumno y por otra parte el adivinar pese menos. Damos por supuesto que todas las *sub-preguntas* (cuatro en el ejemplo de la fig. 3) componen una única pregunta. Más adelante veremos otros métodos de corrección para conseguir lo mismo (aprovechar el conocimiento parcial y reducir la adivinación) aplicables a las preguntas con una única respuesta correcta.

La pregunta entera (el *bloque* compuesto de varias preguntas Verdadero-Falso) puede puntuar más o menos (el máximo sería 1, como suele hacerse en estas preguntas) según el número de respuestas correctas; por ejemplo:

Todas las respuestas correctamente respondidas	:	1.00
Más de la mitad de las respuestas correctamente respondidas	:	0.50
Menos de la mitad de las respuestas correctamente respondidas	:	0.00

El conceder algún tipo de crédito o puntuación a cualquier respuesta Verdadero-Falso correcta (por ejemplo y en el caso de cuatro Verdadero-Falso, 0.25 a cada respuesta correcta), supondría conceder demasiado a preguntas que se pueden acertar fácilmente adivinando. Se puede pensar en claves de corrección similares, que tienen en cuenta *conocimiento parcial,* como ésta (en una pregunta compuesta por cuatro Verdadero-Falso):

Las cuatro respuestas correctamente respondidas	:	1.00
Tres respuestas correctamente respondidas	:	0.67
Dos respuestas correctamente respondidas	:	0.33
Una o ninguna respuesta correctamente respondidas	:	0.00

Los estudios disponibles indican con bastante claridad que la fiabilidad[29] suele ser mayor cuando se utilizan claves que, como las indicadas, tienen en cuenta el *conocimiento parcial.*

Con las claves anteriores el valor de un ítem no es 1 ó 0; puede adquirir diversos valores, entre 0 y 1, según sea el número de respuestas correctas dentro de cada ítem; la fórmula apropiada en el cálculo de la fiabilidad es, en este caso, la del coeficiente □ de Cronbach[30].

La corrección con estas claves, y el cálculo de la fiabilidad y otros análisis, son naturalmente más laboriosos, pero también se pueden utilizar correctoras ópticas y ordenadores con la programación adecuada.

Estas claves de corrección que tienen en cuenta el *conocimiento parcial* y que se pueden emplear cuando hay *varias respuestas correctas* pueden no interesar a veces por otro tipo de razones.

a) Puede sencillamente *no* interesar el *conocimiento parcial*. En algunos casos no saber *algo* puede ser equivalente a *no saber*, por ejemplo cuando las alternativas correctas están relacionadas entre sí y esa relación hay que conocerla; es más, este tipo de formato *(múltiple*

[29] Recordamos que una fiabilidad alta quiere decir que en pruebas semejantes los sujetos hubieran quedado *ordenados* de manera semejante.

[30] La fórmula de Cronbach es realmente la misma que la de Kuder-Richardson 20 empleada habitualmente en los tests objetivos. La única diferencia está en que cuando utilizamos la fórmula de Cronbach los valores de cada pregunta no son dicotómicos (1 ó 0) sino que pueden adquirir varios valores.

Verdadero-Falso) se presta para comprobar conocimiento y comprensión de relaciones entre elementos.

b) Cuando la *ciencia parcial* puede ser peor que la *no ciencia* como reconocer algunos síntomas pero no todos de una enfermedad (como advierten Albanese y Sabers, 1988).

6. Corrección de las pruebas de elección múltiple: problemas del *adivinar* y del *conocimiento parcial* cuando hay una sola respuesta correcta

Dos problemas importantes asociados con los diversos métodos propuestos para corregir y calificar las pruebas objetivas de elección múltiple *con una única respuesta correcta* son:

1º La posibilidad de adivinar la respuesta correcta, por puro azar, y cómo controlar estas respuestas mediante fórmulas correctoras;

2º Cómo utilizar el *conocimiento parcial* del alumno: puede no estar seguro sobre cuál es la respuesta correcta pero sí puede ser capaz de eliminar algunas de las alternativas como claramente falsas.

Los dos problemas, y modos de tratar las respuestas a preguntas objetivas, responden a dos maneras de entender el adivinar en estas situaciones. Cuando el adivinar la respuesta correcta se percibe como un *problema,* se prescinde de que el alumno, cuando escoge una respuesta sin seguridad de que sea la correcta, puede no estar escogiendo una respuesta al azar sino que lo que hace es calcular probabilidades sobre cuál puede ser la repuesta correcta guiándose de su intuición (intuición basada en sospechas fundadas sobre cuál es la respuesta correcta) y de lo que ya sabe sobre la materia. Cuando el problema se sitúa no en el adivinar, sino en tener en cuenta la ciencia parcial y/o insegura del alumno, se acepta de antemano que la dicotomía *saber/no saber* no se da en la práctica (al menos no se da siempre o simplemente puede no darse) cuando se responde a estas preguntas[31].

Tratamos los dos puntos por separado, aunque están relacionados y responden al mismo problema de fondo aunque con distintos enfoques y también, de alguna manera, desde distintas actitudes por parte del profesor. En el segundo punto enunciado, cómo utilizar el conocimiento parcial del alumno, se prescinde de fórmulas correctoras y en cambio se pretende tener en cuenta y aprovechar el conocimiento parcial o inseguro del alumno. Cada enfoque supone un tratamiento metodológico muy distinto, y obviamente supone también que los alumnos reciben instrucciones distintas sobre cómo responder a estas pruebas objetivas.

[31] Posiblemente lo mismo sucede con muchas decisiones de otro orden que hay que tomar en la vida. Una decisión que después se ve que es acertada, no quiere decir que se tomó con absoluta seguridad, de la misma manera que una decisión que luego se juzga como desacertada no quiere decir que se tomó sin fundamento.

6.1. La fórmula de *corrección por adivinación*

Una de las críticas más comunes a las pruebas objetivas de elección múltiple es la posibilidad de *adivinar* la respuesta correcta. Como respuesta a este problema se ha desarrollado una fórmula denominada de *corrección por adivinación* muy utilizada. Esta fórmula no es la única, mencionaremos otras, pero como es la que realmente se ha impuesto es la que vamos a examinar con más detalle. En general las limitaciones de este tipo de correcciones por adivinación se pueden aplicar también a otras fórmulas.

Las ventajas e inconvenientes de esta fórmula han sido objeto de numerosas discusiones y estudios experimentales. Vamos a analizar y a intentar aportar una síntesis de lo mucho que se ha investigado, sin pretender llegar a una respuesta definitiva en un tema controvertido; Lord (1975) expresa bien la falta de unanimidad frente al uso de esta fórmula cuando dice que *religión, política y la fórmula de corrección por adivinación son áreas en las que dos personas bien informadas mantienen con frecuencia posturas opuestas con gran seguridad.*

6.1.1. Qué se presupone en esta fórmula

El nombre de *corrección por adivinación* no es muy afortunado porque parte de un supuesto que no es verdadero necesariamente y que analizaremos con más detalle. Lo que se supone, en la derivación de la fórmula y cuando se aplica, es que cuando un alumno responde a estas preguntas se da una de estas dos situaciones:

a) el alumno conoce la respuesta y responde correctamente,
b) el alumno no conoce la respuesta y en este caso escoge al azar una cualquiera de las alternativas.

En el caso de que el alumno escoja al azar entre, por ejemplo, cuatro alternativas de las que sólo una es correcta, el alumno tiene una probabilidad de acertar y tres de equivocarse. Como consecuencia, y respondiendo al azar, de cada cuatro preguntas respondería correctamente a una y fallaría en tres. En esta fórmula se hace además la difícil suposición de que todas las posibles respuestas son igualmente atractivas para el alumno que ignora la respuesta correcta.

Para eliminar de la puntuación total las respuestas correctas *adivinadas*[32], según los presupuestos anteriores, habría en este caso que restar del total una pregunta acertada por cada

[32] El término *adivinar* en una prueba objetiva puede tener el significado coloquial de escoger una respuesta sin seguridad de que sea la correcta, o el más técnico de escoger una respuesta entre varias que tienen idéntica probabilidad de ser correctas. Por el contexto se puede entender en cada caso a qué llamamos adivinar. Los términos en inglés (idioma en el que se han escrito la mayoría de estas investigaciones) para designar adivinar en sentido propio (idéntica probabilidad de todas las

tres falladas. La fórmula derivada de estas suposiciones, y propuesta ya desde hace años (Thurstone, 1919; Holzinger, 1924) es la siguiente:

$$TC = B - \frac{M}{(k-1)}$$

TC = Total Corregido
B = número de ítems Bien respondidos
M = número de ítems Mal respondidos
k = número de alternativas en cada ítem

Si en un examen de 80 ítems con cuatro respuestas cada uno, un alumno responde a todos *al azar*, lo que suponemos es que acertará en la cuarta parte de los ítems, 20 en este caso, y responderá incorrectamente a 60 ítems. Su total corregido sería en este caso igual a 20 - (60/3) = 20 - 20 = 0.

Esta fórmula no supone que todos los ítems tienen un idéntico número de respuestas, aunque en este caso, el más corriente, la fórmula es de aplicación más fácil.

Si los ítems tienen un *número diferente de posibles respuestas* cada ítem puntúa de esta manera:

si la respuesta es *correcta*, el valor del ítem es = 1

si la respuesta es *incorrecta* el valor del ítem es = $\frac{-1}{(k-1)}$

donde k es en este caso el número de alternativas del ítem. Este cálculo es fácilmente programable.

Las preguntas omitidas no se penalizan, por lo que en caso de duda lo más seguro es dejar la pregunta sin respuesta, y así se indica a los alumnos en las instrucciones. El que el omitir la repuesta en caso de duda sea lo más beneficioso para el alumno es cuestionable y lo discutiremos más adelante.

6.1.2. Otras fórmulas para penalizar la adivinación

Nuestro propósito es referirnos sobre todo a la fórmula anterior, la más utilizada y a la que nos referimos habitualmente cuando hablamos de la *fórmula de corrección por adivinación*. No es sin embargo la única fórmula propuesta; otras fórmulas correctoras,

---

respuestas) suelen ser *wild guessing, blind guessing, random guessing* y *pure guessing*.

también pensadas para penalizar respuestas incorrectas, o con más propiedad, para disuadir respuestas al azar, pueden verse en Abu-Sayf (1979)[33] y en Urosa (1995).[34]

De estas otras fórmulas merece la pena reseñar al menos otra más, propuesta inicialmente por Gulliksen (1950) y que luego encontramos en muchos otros autores (como Ebel, 1965[35]; Traxler, 1966; Traub, Hambleton y Sing, 1969; Lord, 1975). De hecho se trata de una fórmula muy poco utilizada pero tiene interés porque puede influir apreciablemente en la actitud del alumno cuando no sabe con certeza cuál es la respuesta correcta. Los presupuestos son en última instancia los mismos; las respuestas equivocadas son intentos fallidos de acertar sin saber, y es preferible no responder cuando no se sabe.

En esta fórmula el total definitivo es el número de respuestas correctas más las omitidas, dando a las omitidas el valor de 1/k (k es el número de respuestas en cada pregunta; si todas las preguntas tienen cuatro respuestas, cada respuesta omitida vale .25).

La fórmula, si todas las preguntas tienen el mismo número de respuestas, es por tanto:

$$TC_2 = B + \frac{O}{k}$$

TC_2 = Total Corregido
B = número de ítems bien respondidos
0 = número de ítems omitidos
k = número de alternativas en cada ítem

En este caso las respuestas incorrectas no son penalizadas; simplemente no cuentan, y son las omitidas las que *valen algo*, por lo que se anima al alumno a omitir la respuesta en caso de duda en vez de arriesgarse a no ganar nada si se equivoca. Se trata de disuadir al alumno de responder al azar cuando realmente no sabe, premiando el silencio: *en vez de castigar el error se premia el silencio.*

Si comparamos esta fórmula con la tradicional, vemos que con esta fórmula:

a) Se obtienen medias algo más altas, aunque ambas fórmulas están linealmente relacionadas (Budescu y Bar-Hillel, 1993). La relación entre ambas puede verse en esta fórmula que también expresa el total corregido que acabamos de ver (B + O/k):

[33] F.K. Abu-Sayf, autor de numerosos artículos e investigaciones sobre las fórmulas correctoras en las pruebas objetivas, ha sido profesor en la Universidad de Kuwait y colaborador habitual en revistas de investigación americanas. Aportamos este dato simplemente porque el autor, buena fuente informativa sobre estos temas, no responde al perfil nacional de la mayoría de los autores citados en este contexto.

[34] Urosa (cap. II, 1995) hace una revisión de todos los procedimientos y fórmulas de corrección que se han propuesto para controlar o tener en cuenta la adivinación. El primer autor que se plantea la corrección por adivinación es Thurstone en 1919.

[35] En la bibliografía citamos la versión en español de 1977.

$$TC_2 = \frac{N + (k-1)TC_1}{k}$$

N es el número de ítems, k el número de alternativas y TC_1 es el total corregido según la fórmula tradicional.[36]

A efectos prácticos esto quiere decir que las notas serían básicamente las mismas con las dos fórmulas si se utiliza un criterio relativo al grupo.

b) Los tests tienden a aumentar la fiabilidad y validez (Sax y Collet, 1968; Traub y Hambleton, 1972) aunque hemos visto pocos estudios comparativos.

c) Parece que los alumnos suelen preferirla (Waters y Waters, 1971) y esto no resulta extraño. Cuando no se sabe la respuesta correcta, se responde al azar y se falla, lo que se percibe con esta fórmula es que se ha *dejado de ganar un premio,* y lo que se percibe con la fórmula tradicional es que se ha *ganado un castigo.*

Posiblemente esta fórmula merece ser investigada más a fondo. Puestos a disuadir respuestas al azar aplicando fórmulas que penalizan respuestas incorrectas, en casi igualdad de condiciones parece preferible alguna fórmula que cree menos tensión y rechazo por parte de los alumnos, y desde este punto de vista esta fórmula, en la que se premia la abstención frente al error, puede ser preferible.

Traxler (1966) aduce ventajas adicionales de esta fórmula aunque son de otro orden:

a) Como las respuestas omitidas suelen ser menos que las respuestas incorrectas y también más fáciles de *contar,* la corrección, si se hace a mano, es más sencilla y rápida;

b) Los mismos alumnos en clase, y antes de entregar el examen, pueden hacer la primera parte de la corrección contando las respuestas omitidas y colocando el número en un lugar previamente señalado; esta contabilidad pueden hacerla dos alumnos en cada test (distintos del autor del examen) para mayor seguridad (el autor piensa en clases normales y en alumnos de primaria o secundaria);

[36]Se muestra con facilidad que las dos expresiones de TC_2 son iguales si en esta última fórmula substituimos N (número de ítems) por B + M + O (suma de todas las respuestas, bien respondidas, mal respondidas y omitidas) y TC_1 por la fórmula tradicional. En este caso la fórmula que muestra la relación podríamos expresarla así (otra derivación análoga puede verse en Traxler, 1966):

$$TC_2 = \frac{B + M + O + (k-1)\left(B - \frac{M}{k-1}\right)}{k} = \frac{B + M + O + (k-1)B - (k-1)\left(\frac{M}{k-1}\right)}{k} =$$

$$= \frac{B + M + O + (k-1)B - M}{k} = \frac{B + M + O + kB - B - M}{k} = \frac{O + kB}{k} = \frac{O}{k} + \frac{kB}{k} = B + \frac{O}{k}$$

c) Si la corrección se hace por ordenador o métodos mecánicos, puede haber una última opción en cada respuesta para indicar que la respuesta se ha omitido.

En lo sucesivo nos vamos a referir a la fórmula tradicional, pues es la que realmente se ha impuesto e investigado.

6.1.3. Los supuestos de la fórmula son falsos

La suposición que justifica esta fórmula, el alumno o sabe con certeza la respuesta correcta o responde al azar, ha sido muy discutido (puede verse por ejemplo Budescu y Bar-Hillel, 1993).

La mayoría de los alumnos no responden *a ciegas*, sino que suele haber un proceso previo de eliminación de alternativas más probablemente falsas; tampoco *se adivina* en todos los ítems *por igual* porque en los más difíciles se *adivina más,* y tampoco todos los alumnos responden con la misma *estrategia;* son los alumnos que menos saben los que tienden a adivinar más, como ya señalaba Ebel (1968) hace años.

Cuando se ignora la respuesta correcta, las respuestas no se escogen al azar. Los estudios hechos sobre cómo se distribuyen las respuestas incorrectas en cada ítem muestran que estas respuestas no se reparten por igual; es decir cada respuesta incorrecta no es una elección al azar poco afortunada (estudios citados por Budescu y Bar-Hillel, 1993). Las respuestas no se escogen aleatoriamente, sino además del adivinar inteligente que trataremos enseguida, algunos alumnos tienen su propio estilo de adivinar (escoger la primera respuesta, o la segunda, o la última…).

La fórmula no penaliza propiamente los efectos del azar sino las respuestas incorrectas porque se supone que los que menos saben tienden a adivinar más; además se puede responder al azar y acertar. Influyen también factores de personalidad, y los alumnos *más cautos* son los que realmente pueden quedar penalizados.

Esta fórmula no hace que los resultados sean *más genuinos;* no se puede afirmar que los totales *corregidos* representan lo que uno *sabe* una vez eliminados los efectos del azar (Frary, 1988).

6.1.4. El conocimiento parcial: implicaciones de las instrucciones que se dan a los alumnos

No se puede tratar sobre esta fórmula sin tratar al mismo tiempo sobre las instrucciones que se dan a los alumnos para responder. Obviamente si se va a aplicar alguna fórmula correctora, los alumnos deben saberlo.

Las instrucciones que se dan a los alumnos (y que deben darse siempre, al menos oralmente, e incluso por escrito) son importantes porque van a condicionar sus estrategias de respuesta y sus posibilidades de éxito.

Antes de comenzar a responder los alumnos deben saber si va aplicar o no esta fórmula, y en qué medida les puede perjudicar el adivinar o el responder con cierta duda.

Si se va aplicar esta fórmula, que penaliza las respuestas incorrectas, no es lo mismo comunicar a los alumnos una u otra de estas tres orientaciones en las que se sugieren estrategias distintas para responder:

1º En caso de duda es preferible abstenerse porque los errores se penalizan

2º En caso de duda es preferible escoger la respuesta más probable si se puede eliminar con certeza alguna alternativa falsa; si no se puede eliminar ninguna, es preferible abstenerse y no responder.

3º En caso de duda es preferible escoger la repuesta más probable a) aunque no se pueda excluir ninguna respuesta como falsa si b) se tiene un buen conocimiento de la materia, porque en ese caso es más probable responder correctamente

Con estas tres *orientaciones tipo* no pretendemos agotar todos los consejos que se pueden dar a los alumnos para responder (pueden depender de cada profesor), pero sí representan tres orientaciones suficientemente distintas y que pueden condicionar las estrategias de los alumnos frente a la duda cuando no saben con certeza cuál es la respuesta correcta.

Aplicando la fórmula de corrección, y con mayor razón si se advierte al alumno que es preferible omitir la respuesta en caso de duda, se prescinde del *conocimiento parcial* que puede tener el alumno como reconocen y discuten muchos autores (Abu-Sayf, 1975; Ebel, 1977, Budescu y Bar-Hillel, 1993, y muchos otros). Por conocimiento parcial entendemos exactamente eso: un alumno puede saber al menos que alguna respuesta es incorrecta aunque

no sepa con certeza cuál es la correcta, o puede tener un conocimiento *inseguro* y percibe unas respuestas como más probablemente correctas que otras.

Cuando la fórmula se introdujo, en las instrucciones que se daban a los alumnos (y que se siguen dando con frecuencia) se les aconsejaba *no adivinar* (es preferible abstenerse) en caso de duda. Es la primera orientación indicada antes.

Estas instrucciones, todavía habituales, *(es preferible abstenerse en caso de duda)* han sido frecuentemente criticadas porque cuando el alumno elimina las respuestas claramente incorrectas y escoge la respuesta *más probablemente correcta,* tiene más probabilidades de acertar que de equivocarse. Hay estudios que muestran que cuando los alumnos reconsideran las preguntas inicialmente omitidas y escogen *lo más probable*, los aciertos son más de los que se podrían esperar por puro azar (Cross y Frary, 1977; Frary, 1988; Frary, 1989; Albanese, 1986, 1988 y otros autores). Tampoco se trata de una generalización absoluta; otras investigaciones concluyen que esta fórmula no penaliza a ningún grupo particular de alumnos en el sentido de que cuando los alumnos responden de nuevo a las preguntas inicialmente omitidas, su total esperado prácticamente no cambia, aunque en estos casos se trata de muestras de alumnos capaces y especialmente preparados para responder a este tipo de tests (Angoff y Schrader, 1984, 1986).

Las instrucciones *(omitir en caso de duda)* que suelen darse pueden perjudicar a los alumnos que *saben pero no están seguros* o que tienen información *parcial pero útil*. Un estudio de Angoff (con dos muestras de alumnos universitarios de 5000 sujetos cada una) muestra que las ventajas del *adivinar* cuando se va aplicar la fórmula correctora dependen de la habilidad de los sujetos. *Los más capaces y mejor preparados tienden a tener mejores resultados cuando adivinan,* en cambio los menos capaces tienden a tener peores resultados (Angoff, 1989). El adivinar no tiene las mismas ventajas para todos.

En conjunto parece que con las normas que se dan cuando se va a aplicar la fórmula de corrección por adivinación *(es preferible abstenerse en caso de duda, los errores se penalizan)* se penaliza a los alumnos más cautos y sobre todo a los mejor preparados. Estos alumnos, con otras instrucciones *(en caso de duda escoge la respuesta más probable, no se penalizan los errores)*, hubieran tenido probablemente un mejor resultado. Cuando un alumno puede eliminar algunas alternativas, lo más ventajoso para él puede ser arriesgarse y escoger lo más probable.

Cuando se empezó a caer en la cuenta de que a veces el adivinar favorece al alumno, cuando procura adivinar *inteligentemente* buscando lo *más probable*, se empezó a instruir a los alumnos (al menos en muchos casos), para escoger la respuesta *más probablemente*

correcta si podían eliminar con certeza alguna respuesta incorrecta. Las primeras instrucciones en las que se anima a los alumnos a eliminar lo que les parezca *claramente* incorrecto y *adivinar* entre las respuestas que queden, son de Davis (1967). Es la segunda orientación (o *instrucciones*, con más propiedad) que hemos señalado.

Estas instrucciones indican que es preferible omitir la respuesta cuando no se sabe lo suficiente como para eliminar alguna alternativa. Esta norma también ha sido discutida (Budescu y Bar-Hillel, 1993). No es lo mismo *abstenerse si no se puede eliminar ninguna respuesta incorrecta* que *en caso de duda escoger lo más probable*, aunque todas las respuestas ofrezcan alguna posibilidad de ser correctas. Aunque el alumno no pueda descartar ninguna alternativa con seguridad, puede haber alguna percibida simplemente como más probable.

Si lo que se pretende es tener en cuenta el conocimiento parcial o inseguro del alumno (porque en definitiva es beneficioso para él, y darle normas que no tienen en cuenta este conocimiento parcial es perjudicarle) hay que pensar que no todo conocimiento parcial toma la forma de eliminar con seguridad algunas alternativas. Cualquier apreciación *no uniforme* sobre la probabilidad de que una respuesta sea correcta (por ejemplo 40% 20% 20% 20%) es un ejemplo de conocimiento parcial real o tal como lo percibe el propio sujeto. Si las probabilidades son las indicadas (y no 25% 25% 25% 25% que indicarían total ignorancia), el sujeto estaría inclinado a escoger la primera (un 40% de probabilidades percibidas de acertar), pero sin seguridad como para excluir las otras respuestas.

Budescu y Bar-Hillel (1993) cuestionan las instrucciones de no responder si no se puede eliminar ninguna alternativa con certeza porque si responden *lo más probable*, en general tienen las de ganar. Y esto a pesar de que el sujeto medio parece más bien optimista: las respuestas que se responden correctamente suelen ser menos de las que el alumno cree que ha respondido correctamente (estudios citados por Budescu y Bar-Hillel, 1993). La certeza subjetiva no se corresponde con la ciencia objetiva. Además los estudios vistos muestran que el alumno medio no tiene mucha habilidad para distinguir entre *niveles de incertidumbre*.

Sin embargo el animar a los alumnos a responder cuando no conocen la respuesta correcta con certeza y tampoco pueden eliminar ninguna con seguridad, basándose en sus apreciaciones subjetivas sobre las probabilidades que tiene cada repuesta de ser la correcta, puede presentar un problema ético, sobre todo cuando las preguntas están puestas para *atrapar* deliberadamente a los alumnos que tienen un conocimiento insuficiente. Un ejemplo de preguntas de este tipo sería ésta: *¿Cuál de estas tres ciudades está situada más al norte? Nueva York, Madrid, Roma*… Un alumno que no conoce la respuesta correcta puede saber

(conocimiento parcial) que en Nueva York hace más frío que en Roma, y eliminar Roma, que es la respuesta correcta. Indudablemente desde una conciencia ética se puede cuestionar el dar instrucciones u orientaciones a los alumnos a sabiendas de que van a ser perjudiciales para ellos.

Budescu y Bar-Hillel (1993) concluyen que un *adivinar inteligente* puede ayudar a los más capaces, pero perjudicar a los menos capaces, y lo mismo piensa Angoff (1989), para quien se debe animar a los alumnos a escoger la respuesta más probable solamente si pueden eliminar con seguridad una o dos alternativas, pero avisándoles con claridad que deben estar seguros de que su información parcial es *realmente una información válida*. La verdad es que el conocimiento parcial puede ser un conocimiento incorrecto, por lo que la recomendación de Angoff puede no tener mucho sentido en la práctica. Son muchos los sujetos que no *calibran* bien las probabilidades de que una respuesta sea correcta.

Lo que suponen las instrucciones que se dan a los alumnos es que el examinado medio es el examinado ideal que busca la mejor y más racional estrategia para responder correctamente. Se supone con mucha facilidad que el alumno, si puede eliminar algunas alternativas, escogerá la más probable. Sin embargo aquí entran en juego otras variables.

Para concretar este punto: aunque no hay una *conclusión definitiva* sobre qué estrategia se puede y debe recomendar a los alumnos cuando se va a aplicar la fórmula de corrección, sí parece claro que se puede invitar a un *adivinar inteligente* a los alumnos que tienen conciencia de saber y de ir bien preparados al examen; los que en conjunto van mal suelen salir ganando si se abstienen en caso de duda.

6.1.5. Uso de la fórmula y actitudes del profesor

En el fondo en el utilizar o no utilizar esta fórmula subyace una actitud del profesor. Hay profesores a quienes les *aterra* que el que no sabe pueda acertar algunas preguntas adivinando; en cambio no les preocupa que los alumnos que realmente estudian y saben queden peor en un examen por miedo a quedar perjudicados si adivinan (cuando estos alumnos son precisamente los que tienen más probabilidad de acertar, porque *sí saben* aunque no estén seguros). Salvando las distancias, son actitudes que recuerdan a las de la película *doce hombres sin piedad,* en la que se opone el miedo a condenar a un posible inocente frente al miedo de dejar libre a un posible culpable.

Las actitudes del profesor también entran en otro sentido. Los partidarios de la fórmula se preocupan por los *errores aleatorios:* si no se penalizan de alguna manera las respuestas incorrectas, algunos alumnos obtendrán totales más altos simplemente porque *tienen suerte;*

con el uso de la fórmula se pretende evitar un *adivinar ciego* y el que la pura suerte favorezca al que realmente no sabe.

Los que se oponen a la fórmula pretenden evitar *errores sistemáticos*: los alumnos más *cautos* omiten preguntas de las que tienen *conocimiento parcial* o *inseguro* y que hubieran respondido correctamente si no se aplican fórmulas correctoras y se les dan otras normas *(en caso de duda elimina lo claramente falso y escoge lo más probable, sólo se tendrán en cuenta las respuestas correctas; no se penaliza el error).* No es lo mismo un *adivinar ciego* entre todas las alternativas que un *adivinar informado* después de eliminar algunas *(conocimiento parcial).*

En realidad con buenas preguntas y en número suficiente, es muy difícil alcanzar los niveles mínimos solamente adivinando.

6.1.6. Diferencias en los alumnos en su actitud hacia el riesgo

Cuando se omite un ítem, su valor es siempre el mismo, 0, no cuenta en la suma total. Cuando se responde *adivinando,* aunque sea inteligentemente y con conocimiento parcial, las respuestas no valen lo mismo: pueden valer 1 (si de hecho es correcta), ó -1/(k-1) cuando se aplica la fórmula usual de corrección. Es decir, no hay variabilidad en los ítems omitidos, pero sí la hay en los respondidos sin seguridad.

Los sujetos pueden ser distintos en cuanto a sus actitudes hacia el *riesgo*, y de ahí puede depender su estrategia para responder. Los que prefieren la seguridad al riesgo de responder mal, tenderán a omitir en caso de duda. El evitar riesgos (preferir lo seguro, no responder, a un adivinar con cierta probabilidad de salir perjudicado) es una actitud muy común.

La tendencia a omitir está en cierto grado relacionada con el sexo; las mujeres van más *a lo seguro.* Son bastantes los estudios que muestran que los varones aceptan con más facilidad los riesgos del adivinar cuando responden a preguntas objetivas; incluso cuando no se aplica la fórmula de corrección las niñas omiten más ítems cuando precisamente en este caso la estrategia mejor sería responder a todo (este punto está más estudiado en muestras infantiles y adolescentes, Ben-Shakhar y Sinai, 1991, y otros estudios mencionados por Budescu y Bar-Hillel, 1993). Los mismos estudios muestran que esta actitud hacia el riesgo está también relacionada con características sociológicas (en EE.UU. los alumnos de minorías étnicas tienden a adoptar estrategias seguras y a evitar más los riesgos).

En estos estudios (Ben-Shakhar y Sinai, 1991) no se comparan los diversos métodos de corrección de pruebas objetivas y sus respectivas instrucciones (aplicar o no aplicar la

fórmula de corrección), pero sí muestran que el no aplicar la fórmula correctora habitual tampoco es suficiente para eliminar las diferencias en tendencia a adivinar: hay sujetos (en este caso las niñas preferentemente) que sistemáticamente no siguen las instrucciones y no adivinan aunque los errores no se penalicen.

En este estudio los autores utilizan una fórmula para medir la *tendencia a adivinar;* no se trata de una *medida pura* (imposible de construir) pero sí aporta información válida sobre la tendencia a adivinar:

$$\textit{Tendencia a adivinar} = \frac{\text{respuestas adivinadas}}{\text{respuestas adivinadas} + \text{respuestas omitidas}}$$

Y el número de *respuestas adivinadas* (se trata solamente de una estimación) es igual a:

$$\textit{Respuestas adivinadas} = \text{respuestas incorrectas} + \frac{\text{respuestas incorrectas}}{\text{alternativas por pregunta} - 1}$$

6.1.7. Influjo de las características del examen en los riesgos aceptables

Cuando hay un *mínimum* de respuestas correctas para el apto (como sucede en muchos exámenes), el aceptar riesgos depende de cómo siente el alumno que va respondiendo. Si está razonablemente seguro de que ha llegado al mínimum, tenderá a respuestas seguras (omitir en vez de adivinar). En cambio los alumnos que saben que están por debajo del número de respuestas correctas para aprobar, tenderán a aceptar riesgos, e incluso a adivinar ciegamente, sin estrategias determinadas en caso de ignorancia.

Otra situación análoga se da cuando se trata de un examen (un concurso, una oposición) donde el único resultado *premiado* es responder a todo bien. En este caso la mejor estrategia es responder a todas las preguntas, incluso con dudas.

En este tipo de circunstancias, las instrucciones que se dan habitualmente son inapropiadas.

6.1.8. Aplicación de la fórmula y clasificación de los alumnos

Si todos responden a todo y no omiten ningún ítem, y cuando no saben eliminan las respuestas más improbables y escogen al azar entre las que quedan, los resultados de aplicar o no aplicar la fórmula son los mismos con sólo pequeñas diferencias: los alumnos quedan *ordenados* casi de la misma manera con la fórmula y sin la fórmula (Frary, 1988). Por otra parte si los resultados se van utilizar para evaluar o comparar grupos, las consecuencias de utilizar la fórmula de corrección por adivinación son negligibles (Frary, 1989).

6.1.9. Aplicación de la fórmula y fiabilidad del test

Cuando se aplica la fórmula de corrección, la fiabilidad suele ser menor que cuando se contabilizan sin más las respuestas correctas. Los resultados experimentales no son siempre claros, pero en general no se advierte una ventaja *psicométrica* (mayor fiabilidad y también mayor validez) importante con respecto a no utilizar esta fórmula y utilizar sin más el número de respuestas correctas (investigaciones que lo confirman están, entre otras muchas, las de Abu-Sayf, 1975; Cross y Frary, 1977; Rowley y Traub, 1977; Bliss, 1980; Frary, 1982). Este es un punto que se puede comprobar siempre que se desee, y más con la facilidad que supone el uso de programas de ordenador.

6.1.10. Aplicación de la fórmula y tiempo requerido para responder

Otra observación importante (Frary, 1988) es que no se penaliza solamente la *ciencia parcial,* que no se tiene en cuenta, sino a los alumnos que por temperamento tienden a pensar más y a examinar con más detención si tienen base suficiente para eliminar al menos algunas alternativas. Para este tipo de alumnos aumenta la necesidad de *tiempo* y quedan perjudicados a no ser que el tiempo disponible sea más que suficiente. El que, en general, cuando se aplica la fórmula de corrección se tarda más en responder parece confirmado en otros estudios.[37]

6.1.11. En qué circunstancias esta fórmula es más aconsejable

Vamos viendo que esta fórmula de corrección por adivinación presenta aspectos discutidos y discutibles. En algunas circunstancias sin embargo sí parece que se puede recomendar esta fórmula de corrección.

a) Cuando hay poco tiempo para responder; en este caso hay más respuestas al azar y salen ganando los más *audaces.* Sin embargo esta circunstancia, poco tiempo para responder a todas las preguntas, se debe evitar (pensando en el alumno medio, no en el que nunca tendría tiempo suficiente).

b) En tests difíciles (o cuando se presume poca preparación en muchos de los examinados) o cuando los requisitos para el *apto* son más bien bajos, como puede ser el caso en algunas pruebas de selección poco exigentes. En general la fórmula es útil cuando presumiblemente *muchos alumnos* no van a saber responder a *muchas preguntas.*

[37] Budescu y Bar-Hillel (1993) aducen varios estudios al respecto, incluso una tesis doctoral en la que se investiga este punto. No es tan claro que se tarde *mucho más* ni siempre, pero la cuestión *tiempo* es siempre importante en los exámenes.

En exámenes escolares, donde se espera que la mayoría responda a *casi todo*, esta fórmula no tiene ventajas claras con respecto al procedimiento más sencillo de utilizar el número de respuestas correctas (Frary, 1988).

Sí puede ser útil si el examen consta de muy pocas preguntas. Este tipo de pruebas o exámenes suelen constar de muchas preguntas, y en la medida en que hay muchas preguntas el azar en las respuesta va pesando menos. A veces se incluyen unas pocas preguntas objetivas, bien pensadas para comprobar algunos objetivos específicos, en un examen que consta fundamentalmente de preguntas o ejercicios de otro tipo (preguntas abiertas). En muy pocas preguntas las respuestas al azar pueden pesar desproporcionadamente y puede interesar aplicar alguna fórmula disuasoria. Esta situación, pocas preguntas como complemente a otras de otro tipo, no es sin embargo la situación habitual y la que se tiene en cuenta cuando se investiga sobre estas fórmulas.

6.1.12. Consideraciones y conclusiones finales sobre la fórmula de corrección por adivinación

No es fácil hacer una síntesis clara sobre las ventajas e inconvenientes de estas fórmulas, pero sí cabe hacer unas consideraciones finales. Teniendo en cuenta que este tema no se acaba aquí. Hay otros procedimientos muy distintos de los tradicionales, y que veremos en el apartado siguiente, que abordan el problema del *adivinar* y de la *ciencia parcial* con otros procedimientos.

a) Lo mucho que se ha investigado (y se sigue investigando) sobre esta fórmula denominada, posiblemente con poco cierto, de *corrección por adivinación*, muestra que el adivinar en las pruebas objetivas es un problema. Pero este problema es sobre todo del profesor, no del alumno, como bien notan Budescu y Bar-Hillel (1993). Y posiblemente hay problemas mayores para el profesor en estas situaciones de examen (copiar[38], pasar información, conocer de antemano las preguntas, etc.).

b) Sí parece claro que el fundamento de la fórmula no es correcto, la dicotomía saber y responder correctamente o no saber y adivinar, no se da en la práctica. Se trata simplemente de una fórmula disuasoria para evitar respuestas al azar, y si ése es el objetivo, no faltan quienes opinan que la penalización es muy pequeña (Budescu y Bar-Hille, 1993).

[38] Para detectar el *copiar* hay índices que comparan entre sí los exámenes; pueden verse entre otros Frary y Tideman (1997), Sotaridona y Meijer (2003) y Wollack (2003). Este último autor compara los distintos índices, el más recomendable parece ser el denominado índice ω, y advierte que deben utilizarse con muestras no inferiores a 50 sujetos, con sentido común (*wisely*); no para invalidar un examen sino como *evidencia confirmatoria* de otro tipo de información independiente, como es la observación directa del hecho de copiar; tampoco deben suplir las *medidas preventivas* (como es una vigilancia adecuada).

c) Esta penalización 1º no tiene ventajas claras desde un punto de vista psicométrico (fiabilidad, validez); 2º los alumnos quedan *ordenados* más o menos de la misma manera y 3º de hecho se tarda más en responder a todo el examen.

d) Aunque se consiga de hecho disuadir a los propensos a adivinar, a los que no saben, etc., también parece claro que se penaliza a algunos alumnos innecesariamente. Se penaliza a los más lentos, a los que aceptan menos riesgos, se penaliza en función de variables extra-académicas que pueden afectar negativamente a determinados subgrupos de alumnos.

e) Diferentes fórmulas, y sobre todo diferentes instrucciones, inducen a los alumnos a utilizar estrategias diferentes. Es imposible dar instrucciones que sean beneficiosas para todos, pero posiblemente lo adecuado es avisar a los alumnos que si tienen conciencia de saber la materia, en caso de duda tienen más probabilidades de acertar si escogen la respuesta que les parezca más probable, sobre todo si pueden eliminar con bastante seguridad alguna alternativa. En cambio a los menos capaces y preparados les perjudican las respuestas al azar.

f) En cualquier caso sí parece una cuestión de ética profesional (que el profesor puede minimizar, pero que al alumno le importa mucho) el informar a los alumnos con exactitud sobre la fórmula concreta que se va aplicar y cuánto pueden perder con cada error. No se trata de que sepan simplemente que se les aplica una fórmula, sino qué fracción se les resta por cada error (1/k-1 en la fórmula habitual). Muchos alumnos no aplican una estrategia correcta, y salen perjudicados al margen de lo que sepan, por falta de información. Si no informar adecuadamente no parece correcto, mucho menos lo es el dar orientaciones inadecuadas.

g) Los especialistas e investigadores no han llegado a un acuerdo satisfactorio sobre la aplicación de esta fórmula. Una conclusión racional, aunque no la única posible, es prescindir de la fórmula. Se eliminan problemas, y es el único procedimiento que permite dar información y estrategias válidas para todos los alumnos, cualesquiera que sean sus diferencias en preparación, temperamento o motivación. No aplicar la fórmula y animar a todos a responder a todas las preguntas elimina ventajas a los más audaces y no castiga a los más cautos.

h) En este contexto no sobra recordar que podemos determinar con mucha seguridad (con criterios estadísticos) el número máximo de preguntas que un sujeto puede adivinar, e incluso hay tablas ya hechas que tienen en cuenta tanto el número de preguntas como el número de alternativas (Wang y Calhoun, 1997); lo que suele suceder sin embargo es que los profesores, utilizando sus propios criterios, ya suelen poner el apto, de manera espontánea e intuitiva, por encima de lo que es probable acertar adivinando.

6.2. Métodos de corrección que tienen en cuenta el *conocimiento parcial* del alumno

Como hemos visto, los métodos más utilizados para corregir y calificar estas pruebas son dos:

1º utilizar como *puntuación directa* el número de respuestas correctas;
2º aplicar la fórmula de corrección por adivinación.

Otros métodos que se han propuesto buscan *simultáneamente:*

a) reducir o eliminar las respuestas *al azar,*
b) tener en cuenta el *conocimiento parcial* del alumno.

En estos métodos las instrucciones que se dan a los alumnos para responder son distintas; se les da una de estas dos instrucciones en el caso de que duden cuál es la respuesta correcta:

En caso de duda eliminar las alternativas que a su juicio son *probablemente falsas*
En caso de duda escoger las alternativas que a su juicio son *probablemente verdaderas*.

La clave de corrección se adecua al número de respuestas que el alumno considera *probablemente correctas*.

Estos métodos se basan en el supuesto de que un *adivinar inteligente* está de hecho mostrando ciencia aunque esta ciencia sea *insegura*. Para Billing (1974) que propone una serie de métodos de preguntar y corregir para *medir* el *adivinar inteligente* (y que pueden verse en Heywood, 1977, 1989), el *adivinar inteligente* es una habilidad distinta del *conocimiento* y muy útil en la vida. Tampoco se trata siempre y necesariamente de un adivinar inteligente cuando se responde con estas normas. Lo que se le pide al alumno es que señale todas las respuestas probablemente correctas o que elimine todas las respuestas probablemente incorrectas[39].

Exponemos los dos métodos que hemos visto propuestos y que son equivalentes[40].

[39] Estas instrucciones eliminan de las pruebas objetivas la *odiosidad* que provoca el tener que enfrentarse tantas veces seguidas, una en cada pregunta, con la alternativa *acierto/me equivoco*, alternativa *maniquea* que tampoco responde a muchas de las decisiones de la vida. El alumno tiene más margen de maniobra, como sucede en las preguntas de respuesta abierta y posiblemente sus respuestas reflejan más adecuadamente y con más matiz, lo que realmente sabe o no sabe.
[40] Estos métodos pueden verse tratados con amplitud en Urosa (1995).

6.2.1. Eliminar todas las respuestas probablemente falsas

1º En vez de escoger la *respuesta correcta*, al alumno se le instruye para que *elimine todas las alternativas que a su juicio pueden ser falsas* (ver por ejemplo Collet, 1971). La corrección se hace así (k = número de alternativas en cada ítem):

cada eliminación correcta tiene un valor = $\frac{1}{k-1}$

si se elimina la respuesta correcta, el ítem tiene un valor = -1

Por lo tanto **si el ítem tiene:**	**valor de cada *eliminación correcta***
5 alternativas, 1 correcta y 4 falsas:	$\frac{1}{5-1} = 0.25$
4 alternativas, 1 correcta y 3 falsas:	$\frac{1}{4-1} = 0.33$
3 alternativas, 1 correcta y 2 falsas:	$\frac{1}{3-1} = 0.50$
Siempre que se elimine la alternativa correcta el valor del ítem es:	-1
Si se eliminan todas las alternativas falsas, el valor del ítem es:	1

Lo *seguro* para el alumno es eliminar solamente las respuestas falsas de las que está seguro que son falsas. Si se arriesga y elimina la respuesta correcta, tiene mucho que perder. Se elimina la dicotomía a*certar/equivocarse del tod*o (que tampoco es real en la vida ordinaria, fuera de los exámenes) y el alumno puede manifestar lo que sabe, y obtener crédito por ello, aunque su saber sea incompleto. Además se elimina la odiosidad y mal clima que suele generar el uso de la fórmula de corrección por adivinación (y que a fin de cuentas tampoco corrige los efectos del azar).

Collet (1971, y otros autores que cita) encuentra con este sistema una mayor fiabilidad que cuando se aplica la fórmula convencional de corrección por adivinación o se ponderan las respuestas falsas según determinados criterios (y pueden valer 0.50, 0, -0.50 ó -1). La razón parece estar en que se evita mejor el *adivinar* y se tiene en cuenta el *conocimiento parcial* del alumno.

6.2.2. Escoger todas las respuestas probablemente verdaderas

En un método similar al anterior (explicado en Jaradat y Sagawed, 1986), se pide al alumno que en caso de duda *escoja todas las alternativas que a su juicio son probablemente verdaderas*. La corrección es en este caso así:

Si la alternativa correcta:	*valor del ítem*
Está entre las escogidas como probables:	k (número de alternativas) menos número de alternativas elegidas
No está entre las escogidas como probables:	se resta el número de alternativas escogidas

Por ejemplo, y en el caso de 4 alternativas (k = 4),

si el alumno escoge:	*valor del ítem*
4 alternativas, (incluida la correcta necesariamente)	4-4 = 0
3 alternativas, incluida la correcta:	4-3 = 1
2 alternativas, incluida la correcta:	4-2 = 2
1 alternativa, que es la correcta:	4-1 = 3
1 alternativa que *no es* la correcta:	-1
2 alternativas, *sin incluir la correcta:*	-2
3 alternativas, s*in incluir la correcta:*	-3

Con esta clave de corrección la puntuación total máxima sería igual al número total de respuestas falsas en el test. En un test de 10 preguntas, cada una con tres respuestas falsas y una correcta, la puntuación máxima sería 30. Si queremos que la puntuación máxima coincida con el número de ítems, basta dividir los valores anteriores por el número de respuestas falsas en cada ítem.

Jaradat y Sawaged (1986) muestran en su estudio que los coeficientes de fiabilidad y validez (correlación con pruebas equivalentes) son mayores que cuando se aplica la fórmula de corrección por adivinación, o se suman sin más las respuestas correctas, en otras investigaciones se llega a conclusiones parecidas (Urosa, 1995).

De manera semejante a lo que sucede con el método anterior, lo seguro para el alumno es no escoger como respuesta correcta la que probablemente es falsa. El que este procedimiento (escoger lo que probable o ciertamente se piensa que es verdadero) sea preferible al anterior (eliminar lo que probable o ciertamente se juzga que es falso) es algo que se podría investigar.

Estos métodos al menos a) disuaden al alumno de un adivinar ciego, b) por otra parte le animan a mostrar sus conocimientos aunque sean incompletos porque premian la *ciencia parcial* y c) penalizan más que otros sistemas el *falso conocimiento,* tanto si no incluyen la respuesta correcta entre las probablemente correctas como si eliminan la respuesta correcta

entre las probablemente incorrectas (Jennings and Bush, 2006)[41]. La hipótesis que parece probada es que el mejor funcionamiento de estas claves se debe a la disminución de las respuestas al azar y a tener en cuenta la *ciencia parcial* del alumno; los alumnos que saben más o que saben menos, o con mayor o menor seguridad, quedan mejor diferenciados.

6.2.3. Utilización simultánea del método tradicional (una única respuesta correcta) y el de eliminar todas las respuestas probablemente falsas

También se ha propuesto el utilizar estos sistemas simultáneamente con el tradicional (Billing, 1975): el alumno señala en primer lugar la respuesta que cree correcta, y no es penalizado si escoge una alternativa incorrecta, *pero además* elimina las que está seguro que son incorrectas, y en este caso es penalizado si elimina la correcta. Cada ítem tiene por lo tanto dos puntuaciones, y se mantiene la que más favorece al alumno[42]. No hemos visto análisis específicos sobre este sistema doble en el que se pretende distinguir entre *conocimiento seguro* y un *adivinar inteligente*.

Otros procedimientos menos utilizados consisten en *ponderar los ítems* de diversas maneras (la más común es posiblemente utilizar la correlación ítem-total: *pesan más* los ítems más discriminantes y los muy fáciles o muy difíciles pesan menos). En general con este procedimiento aumenta la fiabilidad con respecto a utilizar el número de respuestas correctas o de aplicar la fórmula de corrección por adivinación, pero la diferencia es pequeña si el test es largo; con tests cortos el efecto es mayor (Haladyna, 1985, expone modos de ponderar los ítems y analiza numerosos estudios experimentales).

6.3. Métodos de corrección que tienen en cuenta el *nivel de seguridad* del alumno al responder.

Los sujetos, además de responder a cada ítem, indican su nivel de seguridad en la respuesta; cada ítem por lo tanto tiene una doble respuesta:

a) Escoger la respuesta correcta (que puede ser *verdadero* o *falso*),
b) Indicar el *grado de seguridad* que se tiene en que la respuesta elegida es la correcta.

[41] Estos autores (Jennings y Bush, 2006) examinan desde una perspectiva teórica las diferencias entre *conocimiento perfecto, conocimiento parcial, ausencia de conocimiento, desconocimiento parcial* y *desconocimiento total*, las probabilidades de acertar en cada situación y efectos de estos procedimientos de corrección en todas estas situaciones.

[42] Heywood (1989) explica este sistema en sus diversas modalidades.

Gardner-Medwin (1995, 1998, 2006)[43] es el autor más destacado que ha propuesto y difundido este sistema de corrección, muy utilizado en facultades de medicina inglesas (Gardner-Medwin y Gahan, 2003)[44].

Clave de corrección:

	Grado de confianza en la respuesta correcta:			
	Bajo (1)	Medio (2)	Alto (3)	Sin respuesta
Respuesta correcta	1	2	3	0
Respuesta incorrecta	0	-2	-6	0

Una ponderación negativa tan extrema (0, -2, -6) cuando la respuesta es incorrecta la propone el autor para todo tipo de preguntas aunque inicialmente la reservó solamente para las preguntas *Verdadero-Falso* (en las que la probabilidad de acertar adivinando es mayor), utilizando 0, -1 y -4 para las preguntas de elección múltiple[45].

Este sistema a) se puede aplicar a preguntas de cualquier formato con tal de que la respuesta sea inequívocamente correcta o incorrecta, b) es muy simple y fácil de entender y c) es motivante porque los alumnos salen siempre ganando si son capaces de discriminar correctamente su grado de seguridad y responden sinceramente.

A los alumnos se les explica que para obtener la máxima puntuación por pregunta (nivel 3 de confianza) deben ser capaces de poder justificar su respuesta, hasta el punto de aceptar el riesgo de ser seriamente penalizados si responden incorrectamente. En ejercicios de autoevaluación (y evaluación formativa) los alumnos aprenden a reflexionar sobre su propia seguridad y desarrollan la habilidad de hacer juicios correctos sobre la confianza con que responden.

1) La motivación original para establecer estas claves no fue penalizar respuestas incorrectas, sino *mejorar los hábitos de estudio de los alumnos*, haciéndoles caer en la cuenta de que tanto las respuestas correctas pero inseguras como las respuestas correctas *por casualidad* (adivinando) no equivalen a un conocimiento correcto[46]. Este sistema les invita a reflexionar antes de responder, a buscar relaciones, a pensar en la justificación de su respuesta.

---

[43] El autor ha experimentado sobre todo con estudiantes de medicina; algunas de sus publicaciones sobre *Confidence-Based Marking* pueden verse en su página Web http://www.ucl.ac.uk/~ucgbarg/ en *Teaching Publications.*

[44] No es más utilizado sobre todo por la falta de los programas comerciales adecuados (Gardner-Medwin, 2005).

[45] También se han propuesto otros factores correctores *más benévolos* (1, .30 .10 con signo *más* si la respuesta es correcta y signo *menos* si es incorrecta) (Wilcock, 2004).

[46] Este autor (Gardner-Medwin, 1998) recuerda que el mantener con seguridad lo que es falso como verdadero es especialmente importante en medicina. Por otra parte para el profesor las respuestas incorrectas y dadas con gran seguridad son muy informativas.

Gardner-Medwin (2003) establece esta gradación en el conocimiento que se puede manifestar en la respuestas: *conocimiento correcto y seguro*, *conocimiento correcto pero inseguro, simple ignorancia,* y, lo que es especialmente peligroso, *conocimiento erróneo y seguro* (*misconceptions*)[47]. El objetivo de estas claves de corrección es en definitiva mejorar el estudio de los alumnos; estimularles para que estudien reflexionando, para que tomen conciencia de su propia seguridad y la expresen con sinceridad. El tomar conciencia del grado de seguridad es en sí mismo importante; la inseguridad, por ejemplo en la comprensión de conceptos importantes, puede llevar a ulteriores fracasos.

2) La gradación de las ponderaciones en las respuestas incorrectas (de 0, -2 y -6, que puede parecer excesiva) es importante para motivar respuestas sinceras cuando el alumno manifiesta su grado de confianza. Los buenos alumnos responden correctamente con gran seguridad, mientras que los que no son tan buenos son más prudentes al expresar su grado de seguridad, ya que arriesgan mucho si responden mal y con seguridad. Un aprendizaje importante es el caer en la cuenta de que uno puede ser premiado por reconocer un bajo nivel de confianza. La expresión honesta y correcta del nivel de confianza o seguridad en lo que se dice es un valor en sí mismo y en cualquier contexto comunicativo.

3) La tarea de redactar ítems queda simplificada porque el hacer preguntas complejas es aquí menos importante; preguntas sencillas y directas discriminan mejor que cuando se utilizan las claves convencionales (Gardner-Medwin, 1995). En las preguntas del tipo *Verdadero-Falso* se elimina en buena medida el problema de la adivinación y quedan revalorizadas con estos sistemas de corrección.

4) No hay evidencia de las diferencias entre los sexos que muestran otras investigaciones (Hassmen y Hunt, 1994, Gardner-Medwin, 1995, 2005) (en general las mujeres parecen arriesgarse menos en caso de duda)[48].

5) Las características psicométricas de los tests corregidos estas claves son buenas (mejora la fiabilidad, predice mejor el resultado de otras pruebas que el simple tanto por ciento de respuestas correctas, Gardner-Medwin, 2005).

---

[47] *The original reason for introducing confidence-based testing at UCL (University College of London) was to help students think about and identify where they lie on the scale above, in relation to any and every issue that arises in their studies* (Gardner-Medwin, 1995). *Misconception (uncertain bias towards a wrong answer) about basic issues in a subject can be a huge obstacle when it comes to trying to build higher levels of knowledge, and of course the more confidently the misconceptions are held the worse this can be. So the original rationale was to improve students' study habits* (Gardner-Medwin and Gahan, 2003).

[48] Como hemos comentado en otros lugares, son varios los autores (Hassmen y Hunt, 1994; Ben-Shakhar y Sinai, 1991; Budescu y Bar-Hillel, 1993) que citan numerosas investigaciones en las que se muestra que, en general, las mujeres cambian más sus respuestas que los varones y tienden más a dejar ítems sin responder. Si este sistema elimina diferencias que no tienen que ver con el saber más o menos, ya es un dato a su favor.

6) Los alumnos manifiestan su interés por este sistema cuando se dan cuenta de que les permite ver en qué áreas tienen un conocimiento inadecuado o dónde se están engañando (Gardner-Medwin y Gahan, 2003); también manifiestan que aumenta su autoconfianza y que les fuerza a pensar más. En definitiva este tipo de doble respuesta premia más la comprensión y la capacidad de relacionar que la mera intuición o recuerdo de la respuesta correcta (Gardner-Medwin, 2006).

7) Estas claves de corrección se han utilizado con éxito en evaluaciones formativas (cuya finalidad es *ayudar a aprender*, más énfasis en corregir errores a tiempo que en calificar)[49]. Posiblemente la aplicación *más fácil* de estos sistemas está precisamente en la evaluación formativa, en ejercicios de clase y de autoevaluación, y por supuesto en cursos *online* y en general en la enseñanza asistida por ordenador.

[49] En el University College de Londres (con sistemas programados por ordenador); también se han utilizado en exámenes convencionales de hasta 300 preguntas del tipo *Verdadero-Falso* (Gardner-Medwin y Cahan, 2003).

8. Referencias bibliográficas

1. ABU-SAYF, F.K., (1975). Relative Effectiveness of the Conventional Formula Scoring, *Journal of Educational Research*, 69, 160-162.
2. ABU-SAYF, F.K., (1979). The Scoring of Multiple-Choice Tests: A Closer Look, *Educational Technology*, June, 5-15.
3. ALBANESE, MARK A. A. and SABERS, D. L., (1988). Multiple True-False Items: A Study of Interitem Correlations, Scoring Alternatives and Reliability Estimation, *Journal of Educational Measurement*, 25, 111-123.
4. ALBANESE, MARK A., (1986). The Correction for Guessing: A Further Analysis of Angoff and Schrader, *Journal of Educational Measurement*, 23, 225-236.
5. ALBANESE, MARK A., (1988). The Projected Impact of the Correction for Guessing on Individual Scores, *Journal of Educational Measurement*, 25, 149-157.
6. ALBANESE, MARK A., (1993). Type K and Other Complex Multiple-Choice Items: An Analysis of Research and Item Properties, *Educational Measurement: Issues and Practice*, 12 (1). 28-33.
7. ANGOFF, WILLIAM H. and SCHRADER, W.B., (1984). A Study of the Hypotheses Basic to the Use of Rights and Formula Scores, *Journal of Educational Measurement*, 21, 1-17.
8. ANGOFF, WILLIAM H. and SCHRADER, W.B., (1986). A Rejoinder to Albanese, The Correction for Guessing: A Further Analysis of Angoff and Schrader, *Journal of Educational Measurement*, 23, 237-243.
9. ANGOFF, WILLIAM R., (1989). Does guessing really help? *Journal of Educational Measurement*, 26, 323-336.
10. ARMSTRONG, ANNE-MARIE, (1993). Cognitive-Style Differences in Testing Situations, *Educational Measurement: Issues and Practice*, 12, (3) 17-22.
11. BARANOWSKI, REBECCA A.; DOWNING, STEVEN M.; GROSSO, LOUIS J.; PONIATOWSKI, PAUL A. and NORCINI, JOHN J., (1994). *Item Type and Ability Measured: The Validity of Multiple True-False Items*, paper presented at the annual meeting of the National Council on Measurement in Education, New Orleans.
12. BATEMAN, DIANNE and KATO, CAROLYN, (1993). *The Relationship Between Assessment Tasks and Higher Level Thinking in the Social Sciences*, paper presented at the annual meeting of the American Educational Research Association, Atlanta.
13. BEN-SHAKHAR, G. and SINAI, Y., (1991). Gender differences in multiple-choice tests: The role of differential guessing tendencies, *Journal of Educational Measurement*, 12, 428-454.
14. BILLING, D.E., (1974). The Effect of Guessing on the Results of Objective Tests: A Novel Approach, *Research into Tertiary Science Education*, London, Society for Research into Higher Education.
15. BLISS, L.B., (1980). A Test of Lord's Assumption Regarding Examinee Guessing Behavior on Multiple-Choice Tests Using Elementary School Students, *Journal of Educational Measurement*, 17, 147-152.
16. BLOOM, BENJAMIN S., MADAUS, GEORGE F. AND HASTINGS, J. THOMAS (1981). *Evaluation to Improve Learning.* New York: McGraw-Hill.

17. BRIGHAM YOUNG UNIVERSITY Testing Center. *A detailed guide to help you make better multiple-choice questions* http://testing.byu.edu/info/handbooks.php (consultado 26 Sept. 06)

18. BRUNO, JAMES E. and DIRKWAGER, A., (1995). Determining the Optimal Number of Alternatives to a Multiple-Choice Test Item: An Information Theoretic Perspective, *Educational and Psychological Measurement*, 55, 6, 959-966.

19. BUDESCU, DAVID and BAR-HILLEL, MAYA, (1993). To Guess or not to Guess: A Decision-Theoretic View of Formula Scoring, *Journal of Educational Measurement*, 30, (4), 277-291.

20. BURTON, RICHARD F. (2001). Quantifying the Effects of Chance in Multiple Choice and True/False Tests: question selection and guessing of answers. *Assessment & Evaluation in Higher Education*, 26 (1),41-1.

21. BURTON, STEVEN J.; SUDWEEKS, RICHARD R.; MERRILL, PAUL F. AND WOOD, BUD (1991). *How to Prepare Better Multiple-Choice Test Items: Guidelines for University Faculty*. http://testing.byu.edu/faculty/betteritems.pdf (consultado 26 Sept. 06)

22. CARTER, KATHY (1986). Test-Wiseness for Teachers and Students. *Educational Measurement: Issues and Practice*, 5 (1) 20-23.

23. CIZEK, GREGORY J.; ROBINSON, K. LYNNE and O'DAY, DENIS M (1998). Nonfunctioning Options: a Closer Look, *Educational and Psychological Measurement*, 58 (4), 605-611.

24. COLLET, L.S., (1971). Elimination Scoring: An Empirical Evaluation, *Journal of Educational Measurement*, 8, 209-214.

25. CRONBACH, LEE J., (1941). An experimental comparison of the multiple true-false and multiple-choice test, *Journal of Educational Psychology*, 32, 533-543.

26. CRONBACH, LEE J., (1942). Studies of Acquiescence as a Factor in the True-False and Multiple Choice Items, *Journal of Educational Psychology*, 33, 401-415.

27. CROSS, L.H. and FRARY, R.B., (1977). An Empirical Test of Lord's Theoretical Results Regarding Formula Scoring for Multiple-Choice Tests, *Journal of Educational Measurement*, 14. 313-321.

28. CHREHAN, KEVIN (1989). *The Validity of Two Item-Writing Rules*, paper presented at the annual meeting of the American Educational Research Association, San Francisco.

29. DAVIS, F. B., (1967). A note on the correction for chance success, *Journal of Educational Measurement*, 3, 43-47.

30. DOCHY, FILIP; MENKERKE, GEORGE; DE CORTE, ERIK AND SEGERS, MIEN (2001). The Assessment of Quantitative Problem-Solving Skills with "none of the above"-items. *European Journal of Psychology of Education*, XVI (2) 163-177

31. DOWNING, STEVEN M. (1991). *The Psychometric Effects of Negative Stems, Unfocused Questions, and Heterogeneous Options on NBME Part I and Part II Item Characteristics.* paper presented at the annual meeting of the American Educational Research Association, Chicago.

32. DOWNING, STEVEN M., (1992). True-False, Alternate-Choice, and Multiple-Choice Items, *Educational Measurement: Issues and Practice*, 11 (3), 27-30.

33. DOWNING, STEVEN M., GROSSO, LOUIS J. and NORCINI, JOHN J., (1994). *Multiple True-False Items: Validity in Specialty Certification*, paper presented at the annual meeting of the National Council on Measurement in Education, New Orleans.

34. EAKIN, R.R., (1977). Dodging the Dilemma of True-False Testing, *Educational and Psychological Measurement*, 37, 659-663.

35. EBEL, ROBERT L. (1968). Blind Guessing on Objective Achievement Tests, *Journal of Educational Measurement*, 5. 321-325.

36. EBEL, ROBERT L. (1975). Can Teachers Write Good True-False Items?, *Educational and Psychological Measurement*, 12, 31-35.

37. EBEL, ROBERT L. (1977). *Fundamentos de la medición educacional*, Buenos Aires, Editorial Guadalupe.

38. EBEL, ROBERT L. (1978). The Ineffectiveness of Multiple True-False Test Items, *Educational and Psychological Measurement*, 38, 37-44.

39. EBEL, ROBERT L. (1982). Proposed Solutions to Two Problems of Test Construction, *Journal of Educational Measurement*, 19, 267-278.

40. EBEL, ROBERT L. (1983). The Practical Validation of Tests of Ability, *Educational Measurement: Issues and Practice*, 2 (2), 7-10.

41. EDVARSON, B., (1980). Effect of Reversal of Response Scales in Questionnaires, *Perceptual and Motor Skills*, 50, 1125-1126.

42. FRARY, ROBERT B. (1982). A Simulation Study of Reliability and Validity of Multiple-Choice Test Scores Under Six Response-Scoring Modes, *Journal of Educational Statistics*, 7, 333-351.

43. FRARY, ROBERT B. (1988). Formula Scoring of Multiple-Choice Tests (Correction for Guessing), *Educational Measurement: Issues and Practice*, 7, nº2, 33-37.

44. FRARY, ROBERT B. (1989). The Effect of Inappropriate Omissions on Formula Scores: A Simulation Study, *Journal of Educational Measurement*, 26, 41-53.

45. FRARY, ROBERT B. (1995). More multiple-choice item writing do's and don'ts. *Practical Assessment, Research & Evaluation*, 4(11). http://ericae.net/pare/getvn.asp?v=4&n=11 .

46. FRARY, ROBERT B. and TIDEMAN, NICOLAUS (1997). Comparison of two Indices of Answer Copying and Development of a Spliced Index. *Educational and Psychological Measurement*, 57 (1), 20-32.

47. FRIEDMAN, STEPHEN J and COOK, GREGORY L., (1994). *The Effect of Cognitive Styles on Changing Multiple-choice Answers*, paper presented at the annual meeting of the National Council on Measurement in Education, New Orleans.

48. FRISBIE, DAVID A. (1973). Multiple Choice versus True-False: a Comparison of Reliabilities and Concurrent Validities, *Journal of Educational Measurement*, 10, 297-304.

49. FRISBIE, DAVID A. (1992). The Multiple True-False Item Format: A Status Review, *Educational Measurement: Issues and Practice*, 11 (4), 21-35.

50. FRISBIE, DAVID A. and SWEENEY, D.C., (1982) The Relative Merits of Multiple True-False Achievement Tests, *Journal of Educational Measurement*, 19, 29-35.

51. GARDNER-MEDWIN, A.R. (1995). Confidence Assessment in the Teaching of Basic Science. *Association for Learning Technology Journal.* 3: 80-85. Disponible en http://www.ucl.ac.uk/~ucgbarg/tea/altj.htm (consultado 16, Sept. 06)

52. GARDNER-MEDWIN, A. R. (1998). Updating with Confidence: Do your students know what they don't know?. *Health Informatics*, 4: 45-46. Disponible en http://www.ucl.ac.uk/~ucgbarg/tea/ctilap2.htm (consultado 18 Sept., 06).

53. GARDNER-MEDWIN, A. R. (2005). *Implementing Confidence-Based Marking with your own Technology*. Paper presented at the 12th International Conference of the Association for Learning Technology which was held at the University of Manchester, England (http://www.alt.ac.uk/altc2005/timetable/abstract.php?abstract_id=593, consultado 19 Sept. 06)

54. GARDNER-MEDWIN, A.R. (2006). Confidence-Based Parking-towards deeper learning and better exams. En Bryan C. And Clegg K. (Eds). *Innovative Assessment in Higher Education.* London: Francis and Taylor. Disponible en http://www.ucl.ac.uk/lapt/innovass6.doc (consultado 18 Sept. 06).

55. GARDNER-MEDWIN, A.R. AND GAHAN M. (2003). Formative and Summative Confidence-Based Assessment. Proc. 7th International Computer-Aided Assessment Conference, Loughborough, UK, July 2003, pp. 147-155. Disponible en http://www.ucl.ac.uk/~ucgbarg/tea/caa03.doc (consultado 18, Sept., 06).

56. GLASNAPP, DOUGLAS R. and POGGIO, JOHN P., (1994). *Psychometric Characteristics of the Multiple-Correct Multiple-Choice Items,* paper presented at the annual meeting of the American Educational Research Association, New Orleans.

57. GRAY, GEORGE T. and RACHOR, ROBERT E., (1995). *Do Longer Stems Have Bigger Flowers? An Investigation of Clinically Focused Multiple Choice Items,* San Francisco, paper presented at the annual meeting of the American Educational Research Association.

58. GROSS, LEON J. (1980). *Preparing Examination Items*, National Board of Examiners in Optometry http://www.optometry.org/articles/Item%20Manual%202001.PDF (consultado 26 Sept. 06)

59. GROSSE, M.E. and WRIGHT, B.D., (1985). Validity and Reliability of True-False Items, *Educational and Psychological Measurement,* 45, 1-13.

60. GULLIKSEN, H., (1950). *Theory of Mental Tests*, New York, John Wiley and Sons.

61. HALADYNA, T. M. and DOWNING, S.M., (1985). *A Quantitative Review of Research on Multiple-Choice Item-Writing*, paper presented at the annual meeting of the American Educational Research Association, Chicago.

62. HALADYNA, T. M. and DOWNING, S.M., (1988) *Functional Distractors: Implications for Test-Item and Test Design,* paper presented at the annual meeting of the American Educational Research Association, New Orleans.

63. HALADYNA, T.M. and DOWNING, S. M., (1989). A Taxonomy of Multiple-Choice Item-writing Rules, *Applied Measurement in Education*, 2,(1), 37-50.

64. HALADYNA, T.M., (1985). *A Review of Research on Multiple-Choice Option Weighting,* paper presented at the annual meeting of the National Council on Measurement in Education, Chicago.

65. HASSMEN, PETER and HUNT, DARWIN P., (1994). Human Self-Assessment in Multiple-Choice Testing, *Journal of Educational Measurement,* 31, 149-160.

66. HEYWOOD, JOHN, (1977). *Assessment in Higher Education,* London, John Wiley.

67. HEYWOOD, JOHN, (1989). *Assessment in Higher Education,* 2nd. edit., London, John Wiley,

68. HOLZINGER, K. J., (1924). On scoring multiple-response test. *Journal of Educational Measurement,* 15, 445-447.

69. JARADAT, D. and SAWAGED, S., (1986). The Subset Selection Technique for Multiple Choice Tests: An Empirical Inquiry, *Journal of Educational Measurement,* 23, 369-376.

70. JENNINGS, SYLVIA and BUSH, MARTIN (2006). A Comparison of Conventional and Liberal (Free-Choice) Multiple-Choice Tests. *Practical Assessment, Research & Evaluation*, 11 (8) http://pareonline.net/pdf/v11n8.pdf (consultado 17, Dic., 2006).

71. KEHOE, JERARD (1995). Writing multiple-choice test items. *Practical Assessment, Research & Evaluation*, 4(9). http://ericae.net/pare/getvn.asp?v=4&n=9

72. KNOWLES, SUSAN L. and WELCH, CYNTHIA A. (1992). A Meta-Analytic Review of Item Discrimination and Difficulty in Multiple-Choice Items Using "None-of-the-Above". *Educational and Psychological Measurement*, Vol. 52 (3), 571-577.

73. LARKINS, A.G. and SWINT JR., J.W., (1976) Acquiescence-Dissent Response Set on an Elementary True-False Achievement Test, *Educational and Psychological Measurement,* 36, 1025-1030.

74. LEVINE, M. V. and DRASGOW, F., (1983). The Relation Between Incorrect Option Choice and Estimated Ability, *Educational and Psychological Measurement* 43, 675-685.

75. LORD, FREDERIC M. (1975). Formula Scoring and Number-Right Scoring, *Journal of Educational Measurement,* 12, 7-11.

76. LORD, FREDERIC M. (1977a). Optimal Number of Choices per Item: A Comparison of Four Approaches, *Journal of Educational Measurement,* 14, 33-38.

77. LORD, FREDERIC M. (1977b). Reliability of Multiple-Choice Tests as a Function of Number of Choices per Item, *Journal of Educational Psychology,* 35, 175-180.

78. MARSH, H.W., (1986). The Bias of Negatively Worded Items in Rating Scales for Young Children: A Cognitive-Developmental Phenomenon, *Developmental Psychology*, 22, 37-49.

79. MORALES VALLEJO, PEDRO, (2006). *Medición de actitudes en Psicología y Educación*. 3ª edición revisada. Madrid: Universidad Pontificia Comillas.

80. MORSE, DAVID T. (1998). The Relative Difficulty of Selected Test-Wiseness Skills Among College Students. *Educational and Psychological Measurement,* 58 (3), 399-408.

81. OWEN, STEVE V. and FROMAN, ROBIN D., (1987). What's Wrong With Three-Option Multiple Choice Items? *Educational and Psychological Measurement,* 47 (2), 513-522.

82. PAXTON, MORAG (2000). A Linguistic Perspective on Multiple Choice Questioning. . *Assessment & Evaluation in Higher Education,* 22 (2), 109-111.

83. PETERSON, C.C. and PETERSON, J.L., (1976) Linguistic Determinants of the Difficulty of True-False Test Items, *Educational and Psychological Measurement,* 36, 161-164.

84. POMPLUN, MARK and OMAR, MD HAFIDZ, (1997). Multiple-Mark Items: An Alternative Objective Format? *Educational and Psychological Measurement,* 57 (6), 949-962.

85. RICH, CHARLES E. and JOHANSSON, GEORGE A., (1990). *An Item-level Analysis of None of the Above,* paper presented at the annual meeting of the American Educational Research Association, Boston.

86. ROGERS, W. TODD and HARLEY, DWIGHT (1999). An Empirical Comparison of Three- and Four-Choice Items and Tests: Susceptibility to Test-Wiseness and Internal Consistency Reliability. *Educational and Psychological Measurement,* Vol. 59 (2), 234-247.

87. ROGERS, W. TODD and YANG, PING (1996). *Test-Wiseness:* Its Nature and Application. *European Journal of Psychological Assessment,* Vol. 12 (3), 247-259.

88. RODRIGUEZ, MICHAEL C. (2005). Three Options Are Optimal for Multiple-Choice Ítems: A Meta-Analysis of 80 Years of Research. *Educational Measurement: Issues and Practice,* Vol. 24, nº 2, 3-13

89. ROWLEY, G. L. and TRAUB, R.E., (1977). Formula Scoring, Number-Right Scoring, and Test-Taking Strategy, *Journal of Educational Measurement,* 14, 15-22.

90. SAX, G. and COLLET, L., (1968). The effects of different instructions and guessing formulas on reliability and validity, *Educational and Psychological Measurement,* 28, 1127-1136.

91. SCOULLER, KAREN (1998). The influence of assessment method on students' learning approaches: Multiple choice question examinations versus assignment essay. *Higher Education,* 35, 453-472.

92. SHAHABI, SOHROB and YANG, LIH-MEEI, (1990). *A Comparison Between Two Variations of Multiple-Choice Items and Their Effects on Difficulty and Discrimination Values,* paper presented at the annual meeting of the National Council on Measurement in Education, Boston.

93. SHERMIS, MARK D.; MESS KOCH, CANTAL; PAGE, ELLIS B.; KLEITH, TIMOTHY Z. AND HARRINGTON, SUSANAMARIE (2002). Trait Ratings for Automated Essay Grading. *Educational and Psychological Measurement,* 62 (1), 5-18

94. SOTARIDONA, LEONARDO S. AND MEIJER, ROB R. (2003). Two New Statistics to Detect Answer Copying. *Journal of Educational Measurement,* 40 (1), 53-69

95. STRATON, R.G. and CATTS, R.M., (1980). A Comparison of Two, Three and Four-Choice Items Tests Given a Fixed Total Number of Choices, *Educational and Psychological Measurement,* 40, 357-365.

96. SUBHIYAH, RAJA G. and DOWNING, STEVEN M., (1993). *K-Type and A-Type Items: IRT Comparisons of Psychometric Characteristics in a Certification Examination,* paper presented at the annual meeting of the Annual Meeting of the National Council on Measurement in Education, Atlanta.

97. THURSTONE, LOUIS L., (1919). A method for scoring tests, *Psychological Bulletin,* 16, 235-240.

98. TOLLEFSON, NONA (1987). A Comparison of the Item Difficulty and Item Discrimination of Multiple-Choice Items Using the "None of the Above" and One Correct Response Options, *Educational and Psychological Measurement,* 47 (2), 385-400.

99. TRAUB, R. E., and HAMBLETON, R. K., (1972). The effects of scoring instructions and degree of speedness on the validity and reliability of multiple-choice tests, *Educational and Psychological Measurement*, 32, 737-758.

100. TRAUB, R. E., HAMBLETON, R. K. and SINGH, D., (1969). Effects of promised reward and threatened penalty on performance of a multiple-choice vocabulary test, *Educational and Psychological Measurement*, 29, 847-862.

101. TRAXLER, ARTHUR E., (1966). Administering and Scoring the Objective Test, en LINQUIST, E. F., (Ed.), *Educational Measurement*, Washington, D.C., American Council on Education, 329-416).

102. TREVISAN, MICHAEL S. and SAX, GILBERT, (1990). *Reliability and Validity of Multiple-Choice Examinations as a Function of the Number of Options per Item and Student Ability*, paper presented at the annual meeting of the American Educational Research Association, Boston.

103. TREVISAN, MICHAEL S.; SAX, GILBERT and MICHAEL, WILLIAM B., (1994). Estimating the optimum number of options per item using an incremental option paradigm, *Educational and Psychological Measurement*, 54, 1, 86-91.

104. UNIVERSITY OF MINNESOTA, The Office of Measurement Services (en Classroom Resources) http://oms.umn.edu/oms/index.php (consultado 26 Sept. 06)

105. UROSA SANZ, BELÉN MERCEDES (1995). *La adivinación en las pruebas objetivas: alternativas a la fórmula clásica de corrección*, tesis doctoral, Madrid, Universidad Pontificia Comillas.

106. WANG, JIANJUN and CALHOUN, GEORGE (1997). A Useful Function for Assessing the Effect of Guessing on True-False and Multiple-Choice Tests. *Educational and Psychological Measurement*, 57 (1), 179-185.

107. WASON, P.C., (1961). Response to Affirmative and Negative Binary Responses, *British Journal of Psychology*, 52, 133-142.

108. WATERS, C.W. AND WATERS, L. K., (1971). Validity and likeability for three scoring instructions for a multiple-choice vocabulary test, *Educational and Psychological Measurement*, 31, 935-938.

109. WILCOCK, SEAN (2004) From unconscious to conscious learning using MCQs with a Confidence Factor, Leeds Metropolitan University, School of Information Management, www.lmu.ac.uk/ies/im/RIP2004-2.pdf (Consultado 19, Sept. 06)

110. WITKIN, H. A., OLTMAN, P. K., RASKING, E. and KARP, S. A., (1971). *Manual for the Embedded Figures Test and Group Embedded Figures Test*, Palo Alto, CA: Consulting Psychologist Press.

111. WOLLACK, JAMES A. (2003). Comparison of Answer Copying Indices with Real Data. *Journal of Educational Measurement*, 40 (3), 189-205.