Psicothema 2006. Vol. 18, nº 4, pp. 848-856
www.psicothema.com

ISSN 0214 - 9915 CODEN PSOTEG
Copyright © 2006 Psicothema

Errores de interpretación de los métodos estadísticos: importancia y recomendaciones

Héctor Monterde i Bort, Juan Pascual Llobel y María Dolores Frías Navarro
Universidad de Valencia

Trabajos empíricos previos han identificado las opiniones de los investigadores respecto de las pruebas de significación y de otros recursos estadísticos, algunas de estas opiniones resultan totalmente inaceptables. En esta investigación comprobamos el grado en que estos errores aparecen en una muestra española de profesores e investigadores de universidad mediante la aplicación de un cuestionario. Los datos obtenidos son importantes para: a) la prevención de interpretaciones inadecuadas; b) la corrección de usos incorrectos; c) el análisis de alternativas posibles; y d) proponer cambios editoriales en los criterios de publicación.

Interpretation mistakes in statistical methods: Their importance and some recommendations. Prior research has identified some of the most common misconceptions on how researchers interpret the results of significance tests. In this study, we examine the scope of these misconceptions in a sample of university professors and researchers from Spain on the basis of a short questionnaire. The obtained results provide important information: i) to prevent wrong interpretation from the data, ii) to correct the wrong use of statistical testing, iii) to suggest new ways of examining the data, and iv) to propose some modifications on the editorial criteria to publish scientific work.

Los filósofos de la ciencia interesados por averiguar lo que distingue al conocimiento científico del no científico han propuesto algunos principios metodológicos generales, como el principio de falsación de Popper (1963), a pesar de los múltiples problemas que plantea su aplicación por la indeterminación de las teorías (Harding, 1976). Quizá por ello los psicólogos, con larga tradición de debate acerca del estatus epistemológico de su disciplina, han dirigido la atención al análisis de las prácticas diarias mediante las que se construye, negocia y comunica la investigación científica, entre las que incluimos la instrumentalización utilizada, las técnicas de transformación y análisis de datos o las tecnologías de representación aplicadas en la formulación y debate de los problemas estudiados.

Históricamente, los científicos sociales y especialmente los psicólogos han confiado en la «comprobación de la significación estadística» como la técnica por excelencia del análisis de datos. Como afirma Cohen (1990), el esquema de Fisher es tremendamente competitivo y atractivo a la vez por su ventaja de ofrecer «*a deterministic scheme, mechanical and objective, independent of content and led to clear-cut yes-no decisions*» (p. 1.307). En Psicología ha gozado de gran predicamento y difusión amplia y creciente (Huberty, 1993).

Sin embargo, en las últimas décadas han crecido exponencialmente las publicaciones que critican la aplicación inadecuada de esta estrategia analítica, lo insatisfactoria que resulta para alcanzar el objetivo final de toda ciencia, acumulación de conocimiento, y por la utilización casi exclusiva como único criterio de interpretación de «la significación de los resultados» (Borges, San Luis, Sánchez y Cañadas, 2001; Falk, 1998; Krueger, 2001; Nickerson, 2000). Las críticas han arreciado en Psicología, en Medicina, en Biología y en otras muchas ciencias.

Autores hay que han llegado a emitir calificaciones muy peyorativas sobre las pruebas de significación estadística, normalmente basadas en el contraste de la hipótesis nula (*Null Hipotesis Significance Testing –NHST–*). Baste como ejemplo las recopiladas por Morgan (2003), algunas de las cuales destacan por su radical extremismo: «*mindlessness in the conduct of research*» (Bakan, 1966, p. 436), «*corrupt form of the scientific method*» (Carver, 1978, p. 397), o «*disastrous*» (Schmidt y Hunter, 2002, p. 66), o la de Rozeboom (1997), para quien «*is surely the most bone-headedly misguided procedure ever institutionalized*» (p. 335). Otras no recogidas por este autor merecen ser destacadas también para situar el alcance de la polémica, como la de Bracey (1988, p. 257), «*statistical significance has nothing to do with meaningfulness*», que ahonda en la diferencia entre significación estadística y significación «social» (o práctica), y las de Cox (1977), «*they* (refiriéndose a los tests de significación) *are also widely overused and misused*», y Moore (1992, p. 2), «*test of statistical significance are overused and misused in an attempt to make a poor or mediocre study appear good*», que coinciden en una idea que, en nuestra opinión, es mucho más representativa de la realidad actual.

En parte debido a estas críticas el *APA* creó el 28 de febrero de 1996 el grupo *Task Force on Statistical Inference (TFSI)*, que en su segunda reunión (Wilkinson and The APA Task Force on Statistical Inference, 1999) propuso reformar las prácticas habituales

Fecha recepción: 7-3-06 • Fecha aceptación: 21-7-06
Correspondencia: Héctor Monterde i Bort
Facultad de Psicología
Universidad de Valencia
46010 Valencia (Spain)
E-mail: hector.monterde@uv.es

de los psicólogos en el análisis de datos y en la elaboración de los correspondientes informes. La quinta edición del *APA Publication Manual* (APA, 2001) se hacía eco de algunas de estas reformas. Desgraciadamente, el cumplimiento de las mismas obligaba a superar ciertas inercias del pasado y quizá por ello la efectividad de las propuestas ha sido hasta el día de hoy más bien escasa. Con toda seguridad se requieren políticas editoriales más firmes y nuevos currículos académicos de formación de los futuros investigadores para garantizar un cambio óptimo en la línea prevista.

Este propósito nos llevó a iniciar una línea de investigación desde la Universidad de Valencia, para la que el estudio realizado por Mittag y Thompson (2000), replicado por Gordon (2001), nos pareció un buen punto de partida (Monterde-Bort, Pascual y Frías, 2005).

Tres objetivos básicos nos han guiado:

1) Estimar la prevalencia de falsas creencias e inadecuadas interpretaciones sobre la metodología estadística en la población española de investigadores en el campo de la Psicología y Ciencias Sociales.
2) Realizar una réplica en España de dos estudios equivalentes realizados en Estados Unidos, que permita un estudio comparativo intercultural.
3) Detectar aquellos conceptos metodológico-estadísticos sobre los que existe mayor confusión, lo que nos ayudará a mejorar nuestra labor docente en el campo de la metodología estadística.

Método

Participantes

Para que nuestros datos fueran comparables a los obtenidos por los estudios de Mittag-Thompson (2000) y Gordon (2001) utilizamos una muestra constituida por 119 profesores-investigadores (PDI) pertenecientes a las facultades de Psicología existentes en las universidades españolas. Todos ellos con el grado de Doctor o suficiencia investigadora.

La distribución por sexos fue de 62 hombres (52.1%), 49 mujeres (41.2%) y 8 casos (6.7%) que no informaron de su sexo.

La distribución por áreas de conocimiento fue la siguiente:

- 50 (42.0%) profesores-investigadores de áreas metodológicas (estadística, psicometría, diseños de investigación…).
- 64 (53.8%) profesores-investigadores de otras áreas (clínica, educativa, organizacional…).
- 5 (4.2%) que no dieron esta información.

Reseñar que la población de referencia del estudio de Gordon estaba constituida por los miembros del AVERA (*American Vocational Education Research*), de cuyo directorio se obtuvo la muestra correspondiente, y el estudio de Mittag y Thompson obtuvo los datos de los miembros del AERA (*American Educational Research Association*).

Realizamos un pre-estudio de fiabilidad por el método Test-Retest con una submuestra seleccionada aleatoriamente, obteniendo coeficientes aceptables (en torno a 0,80).

Instrumentos

Para la recogida de información se aplicó el «*Psychometrics Group Instrument*» de Mittag (1999), el mismo que el utilizado en los dos estudios comparados (Mittag y Thompson, 2000; y Gordon, 2001), desarrollado originariamente para determinar las percepciones de los participantes sobre los tests de significación estadística y otras cuestiones estadísticas. Consta de 29 *ítems* a los que se responde usando una escala tipo Likert de 5 puntos (1= «desacuerdo», 2= «algo en desacuerdo», 3= «neutral», 4= «algo de acuerdo», y 5= «acuerdo»).

En el cuestionario original los *ítems* están expresados en forma positiva y negativa para minimizar los efectos de aquiescencia en las respuestas. Por ello, posteriormente a la aplicación y al igual que se hizo en los dos estudios comparados, los 14 *ítems* cuya afirmación se considera falsa fueron recodificados para invertir sus escalas de respuesta, de forma que la puntuación tomada indicara grado de acierto (en lugar de grado de acuerdo con la afirmación expresada en el *ítem*), es decir, grado de conocimiento de los participantes sobre el tema estudiado: 5= «desacuerdo» (mayor grado de acierto),…, 1= «acuerdo» (menor grado de acierto).

A efectos de clarificar mejor los resultados, los 29 ítems del cuestionario fueron agrupados en los 9 factores lógicos propuestos por Mittag y Thompson en su estudio (2000): «*percepciones generales*», «*percepciones sobre el Modelo Lineal General*», «*percepciones sobre los métodos paso-a-paso*», «*percepciones sobre la fiabilidad de las puntuaciones*», «*percepciones sobre los errores tipo I y II*», «*percepciones sobre la influencia del tamaño de las muestras*», «*percepción de las probabilidades estadísticas como una medida del tamaño del efecto*», «*percepción de los valores 'p' como medidas directas de la importancia del resultado*» y «*percepción de los valores 'p' como medidas de la probabilidad de replicación de los resultados (y del error de estimación de los parámetros)*».

Procedimiento

El proceso de reclutamiento fue equivalente al realizado en los estudios de Mittag-Thompson (2000) y Gordon (2001): población constituida por personal docente-investigador en activo durante los dos cursos escolares 2002-2003 y 2003-2004 de las facultades de Psicología existentes en España. El cuestionario fue enviado y devuelto por correo ordinario. En el período de dos años se recogieron 119 cuestionarios válidos, lo que representó un ratio de respuesta en torno al 30% establecido por Kerlinger (1986) [35% en el estudio de Gordon y 21.7% en el de Mittag-Thompson].

Para el análisis de los datos obtenidos a través del cuestionario se utilizó el paquete estadístico SPSS® versión 12 y la hoja de cálculo EXCEL® de Microsoft.

Resultados

Presentaremos los resultados por factores, construidos éstos conforme a la agrupación lógica descrita anteriormente. Cada factor lógico se presentará en una tabla en la que se incluyen los *ítems* que lo componen, precedidos por el número identificativo en la escala original (entre paréntesis el número identificador usado en nuestro estudio) y, como resultados, la media y límites inferior y superior del intervalo de confianza (al 95% calculado sobre el error típico de la media) de los tres estudios comparados. De estos valores, los correspondientes al estudio de Mittag-Thompson (2000) fueron estimados por nosotros, ya que los autores presentaron la información en formato exclusivamente gráfico, y los correspondientes al de Gordon (2001) fueron calculados por noso-

tros, ya que el autor sólo ofreció los valores de medias y desviaciones típicas. Realizados estos cálculos numéricos se remitieron a los autores para su conveniente sanción antes de iniciar su interpretación y posterior publicación.

Los *ítems* considerados falsos (cuyo menor grado de acuerdo –mostrado por los sujetos– resulta en una puntuación final –mostrada en las tablas– más alta en 'conocimientos') son señalados entre comillas.

Los resultados correspondientes al primero de los factores lógicos establecidos aparecen en la tabla 1. El factor está constituido por los primeros 5 *ítems*, en los que se pregunta sobre el conocimiento general que se tiene de ciertos problemas estadísticos y sobre el debate en torno a la comprobación de la significación estadística.

Los resultados en nivel de conocimiento alcanzados por las tres muestras comparadas (las dos de EE.UU. y la española) son en general bastante coincidentes pero con matices. La muestra española sugiere mayor nivel de duda con lo planteado por el *ítem* 3 y mayor nivel de claridad con lo planteado por el *ítem* 5; a saber, la información acerca de la insuficiente potencia de los estudios psicológicos es bastante inferior entre españoles; en cambio, es mejor la comprensión de lo que se entiende por significación estadística.

Es interesante destacar los resultados obtenidos para el *ítem* 4. Se comprueba que las tres muestras son altamente coincidentes en mostrar bastante *desacuerdo* con la necesidad o conveniencia de que desaparezcan las pruebas de significación estadística, siendo la española, a tenor de su puntuación, la más reacia al cambio. (¿Hay un desconocimiento general de debate? ¿Se le considera un asunto menor? ¿Implica ello el desconocimiento de las recomendaciones del APA? Cuestiones abiertas y sugerentes pero sin respuesta clara).

En segundo lugar ofrecemos la tabla comparativa de resultados correspondiente al segundo de los factores lógicos (véase tabla 2).

Tabla 1
Factor I Percepciones generales

	Mittag y Thompson (2000)			Gordon (2001)			Nuestro (muestra de profesores Univ.)		
	Media (*)	**Límite inf.(*)**	**Límite sup.(*)**	**Media**	**Límite inf.**	**Límite sup.**	**Media**	**Límite inf.**	**Límite sup.**
1(21) Las controversias sobre el uso de los pruebas de significación estadística existen desde hace mucho tiempo y, con toda seguridad, seguirán todavía muchos años más	4,38	4,10	4,65	4,47	4,28	4,66	3,53	3,34	3,72
2(22) Sería mejor usar la frase «estadísticamente significativo» en lugar de «significativo» para describir los resultados en los que la hipótesis nula es rechazada	4,51	4,22	4,80	4,25	3,98	4,52	4,36	4,18	4,54
3(23) La mayoría de los estudios de investigación tienen insuficiente potencia estadística para evitar el error Tipo II	3,56	3,30	3,82	3,41	3,15	3,67	2,88	2,67	3,09
4(24) La ciencia progresaría más rápidamente si las pruebas de significación estadística fueran suprimidas de los artículos publicados	1,70	1,50	1,90	1,70	1,25	2,15	1,59	1,43	1,75
5(25) La significación estadística informa que el investigador rechazó la hipótesis nula.	2,99	2,75	3,22	3,02	2,57	3,47	3,71	3,43	3,99

Notas:
– (*) valores estimados
– Mayor valor (media y límites) significa mayor grado de acierto
– Entre comillas: ítems considerados falsos (escala resp. fue invertida)

Tabla 2
Factor II Percepciones sobre el modelo Lineal General (MLG)

	Mittag y Thompson (2000)			Gordon (2001)			Nuestro (muestra de profesores Univ.)		
	Media (*)	**Límite inf.(*)**	**Límite sup.(*)**	**Media**	**Límite inf.**	**Límite sup.**	**Media**	**Límite inf.**	**Límite sup.**
12(32) Todos los análisis estadísticos («t» de Student, «r» de Pearson, ANOVA…) son correlacionales	2,83	2,55	3,10	2,37	2,01	2,73	2,00	1,76	2,24
26(46) «No es posible usar regresión para comprobar estadísticamente la hipótesis de nula de que las medias de diferentes grupos son iguales»	3,83	3,55	4,10	3,70	3,43	3,97	3,29	3,00	3,58

Notas:
– (*) valores estimados
– Mayor valor (media y límites) significa mayor grado de acierto
– Entre comillas: ítems considerados falsos (escala resp. fue invertida)

Está constituido por 2 *ítems* que abordan creencias sobre el Modelo Lineal General (M.L.G.), uno de los *ítems* puntuado como 'verdadero', el 12, y otro, el 26, como 'falso', criterios de verdad/falsedad basados en la argumentación establecida desde los años sesenta por estadísticos como Cohen (1968) de que todas las variantes de la estadística paramétrica constituyen una sola familia y son fundamentalmente correlacionales. Así, hoy es ampliamente aceptado entre los metodólogos que la regresión/correlación y el análisis de la varianza comparten el mismo modelo matemático, el M.L.G.

La muestra española se distancia de las norteamericanas, demostrando un mayor desconocimiento del M.L.G. y congruente, esa es nuestra presunción, con un modo de entender y enseñar la estadística que algunos autores han definido como el modelo de «*caja de herramientas*», donde la programación didáctica va directamente dirigida a describir el catálogo de técnicas estadísticas posibles según circunstancias, y menos unitario, integrador y matemáticamente fundamentado como el que aporta el M.L.G. Esta constatación es más importante de lo que parece, pues justifica muchos prejuicios carentes de significado, algunos de los cuales quedan al descubierto en estas palabras: «*Some disciplines within the behavioral sciences (e.g. experimental psychology) have had a misperception that MRC (Multiple Regression/Correlation) is only suitable for non-experimental research. We considerer how this misperception arose historically, note that MRC yield identical statistical tests to those provided by ANOVA yet additionally pro - vides several useful measures of size of the effects*»…«*MRC, ANO - VA and ANCOVA are each special cases of the general linear mo - del in mathematical statistics*» (Cohen, West, Cohen y Aiken, 2003, pp. 3-4).

La siguiente tabla corresponde al tercero de los factores lógicos (véase tabla 3). Éste está formado también por dos *ítems* que indagan sobre las creencias relacionadas con la utilidad de los métodos paso-a-paso (*stepwise*) asociados a los procedimientos multivariados. Ambos *ítems* son puntuados como 'falsos', pues denotan falsas creencias bastante comunes entre los investigadores.

De los resultados comparados se confirma la importancia que tienen estas falsas creencias entre la población de los investigado - res sociales, si bien la muestra utilizada por Gordon presenta un mayor conocimiento sobre estos métodos. No es buena noticia que la muestra española se posicione en la zona neutra, cuando los ítems incorrectos demandan mayor rechazo, puesto que los métodos *stepwise* han sido duramente criticados en su capacidad por identificar correctamente al mejor conjunto de predictores para un determinado número de variables predictoras (Thompson, 1998).

La siguiente tabla corresponde al cuarto de los factores lógicos (véase tabla 4). Éste está formado por cuatro *ítems* que abordan un tema psicométrico, el de la fiabilidad de las puntuaciones, expresando creencias fuertemente arraigadas entre los profesionales e investigadores de Psicología. Los tres primeros *ítems* puntuados como 'verdaderos' y el cuarto como 'falso'. Especialmente el primer *ítem* de este grupo, el número 7, puede llamar la atención por su contraste con la tradición psicométrica; según los criterios actualmente adoptados por la APA, como resultado de una también persistente corriente crítica, es que un instrumento psicológico o test no es *fiable* o *infiable* en sí, sino que la fiabilidad es una propiedad de las puntuaciones de un test para una población particular de examinados (Wilkinson and The APA Task Force on Statistical Inference, 1999, p. 596).

En consecuencia con lo dicho, el *ítem* 7, tal y como está formulado, debería ser mayormente aceptado por los sujetos examinados. Los resultados comparativos parecen situar a la muestra española como mejor conocedora del significado (moderno) del concepto «fiabilidad» de un test.

Respecto a los dos últimos *ítems*, que abordan la relación entre la fiabilidad de los datos y la comprobación/estimación de efectos, las mayores distancias se producen con la muestra utilizada por Mittag y Thompson, que obtiene menor nivel de acierto, siendo la española bastante coincidente con la de Gordon. En este factor la puntuación de la muestra española es superior en todos los casos, lo que debe verse como algo positivo a todas luces.

La siguiente tabla corresponde al quinto de los factores lógicos (véase tabla 5). Éste está formado por cuatro *ítems* que evalúan conocimientos asociados a los errores típicos que pueden cometerse en las pruebas de significación estadística, errores tipo-I y tipo-II. Uno de los *ítems*, el primero, es puntuado como 'verdadero' y los tres restantes como 'falsos', los criterios son bastante evidentes y cobran importancia si atendemos al uso ordinario que se hace de las pruebas estadísticas de significación.

Con excepción del *ítem 9*, cuyo nivel de «desconocimiento» es equivalente en las tres muestras, la muestra española está bien informada respecto de estos aspectos comparada con las muestras

Tabla 3
Factor III Percepciones sobre los procedimientos de análisis paso a paso

	Mittag y Thompson (2000)			**Gordon (2001)**			**Nuestro (muestra de profesores Univ.)**		
	Media (*)	**Límite inf.(*)**	**Límite sup.(*)**	**Media**	**Límite inf.**	**Límite sup.**	**Media**	**Límite inf.**	**Límite sup.**
13(33) «En regresión y otros análisis, el método «paso a paso» (stepwise) puede razonablemente ser usado para identificar el mejor subgrupo de predictores de entre un grupo dado»	2,58	2,25	2,90	3,55	3,26	3,84	2,47	2,26	2,68
20(40) «Cuando los investigadores utilizan el método de análisis «paso a paso» (stepwise), el orden de entrada de las variables constituye un indicador muy útil de la importancia de cada variable introducida»	2,99	2,75	3,22	3,47	3,17	3,77	2,72	2,47	2,97

Notas:
– (*) valores estimados
– Mayor valor (media y límites) significa mayor grado de acierto
– Entre comillas: ítems considerados falsos (escala resp. fue invertida)

estadounidenses, que además ofrecen resultados poco consistentes entre ellas.

La siguiente tabla corresponde al sexto de los factores lógicos (véase tabla 6). Está formado por tres *ítems* que intentan comprobar hasta qué punto se tiene conocimiento del efecto o influencia que tiene el tamaño de la muestra sobre los resultados de la significación estadística, sabiendo que el nivel de significación está determinado por el producto «*tamaño del efecto* × *tamaño de la muestra*». Los tres *ítems* son puntuados como 'verdaderos', por lo que se espera que estén más de acuerdo con los enunciados aquellos examinados que posean mayor conocimiento sobre esta circunstancia.

Aquí los resultados de la comparación intercultural son dispares. La muestra española obtiene mayor grado de acierto (y en este caso de 'acuerdo') con lo planteado por el primero de los *ítems*, el 10, sin embargo, falla más que las estadounidenses en los otros dos *ítems*, que son precisamente, a nuestro juicio, los que inciden realmente en el problema. Aunque la significación estadística es dependiente del tamaño de la muestra, esta relación es insuficientemente entendida por la muestra española.

Finalmente, los tres últimos factores lógicos, con tres *ítems* cada uno, abordan confusiones y/o falsas creencias relacionadas con la interpretación de la probabilidad asociada a las pruebas de significación estadística, el valor *'p'*: como medida del efecto, como medida de la importancia del resultado, y como medida de replicabilidad de los resultados (y/o del error de estimación de los parámetros).

La tabla 7 corresponde al séptimo de los factores lógicos. Tres *ítems* que indagan el conocimiento sobre la relación entre el valor *'p'* y el efecto, dos formulados como 'falsos' y uno, el 24, como 'verdadero'.

Tabla 4
Factor IV Percepciones sobre la fiabilidad de las puntuaciones

	Mittag y Thompson (2000)			Gordon (2001)			Nuestro (muestra de profesores Univ.)		
	Media (*)	**Límite inf.(*)**	**Límite sup.(*)**	**Media**	**Límite inf.**	**Límite sup.**	**Media**	**Límite inf.**	**Límite sup.**
7(27) La expresión «la Fiabilidad del Test» constituye una falsedad, ya que la fiabilidad no es una característica de ningún test en sí	2,80	2,50	3,10	2,85	2,48	3,22	3,49	3,21	3,77
19(39) Comprobar la significación de un coeficiente de fiabilidad o de validez con r^2= 0 como hipótesis nula no es útil ni productivo	2,85	2,60	3,10	2,80	2,49	3,11	2,98	2,70	3,26
23(43) La utilización de datos poco fiables provoca una disminución o atenuación de los efectos que van a ser estadísticamente comprobados	2,60	2,30	2,90	3,62	3,27	3,97	3,42	3,16	3,68
28(48) «La fiabilidad no afecta directamente a la probabilidad de obtener significación en un estudio concreto»	2,63	2,35	2,90	3,45	3,08	3,82	3,16	2,90	3,42

Notas:
– (*) valores estimados
– Mayor valor (media y límites) significa mayor grado de acierto
– Entre comillas: ítems considerados falsos (escala resp. fue invertida)

Tabla 5
Factor V Percepciones sobre errores Tipo I y Tipo II

	Mittag y Thompson (2000)			Gordon (2001)			Nuestro (muestra de profesores Univ.)		
	Media (*)	**Límite inf.(*)**	**Límite sup.(*)**	**Media**	**Límite inf.**	**Límite sup.**	**Media**	**Límite inf.**	**Límite sup.**
9(29) Cometer error Tipo II es imposible cuando los resultados obtenidos son estadísticamente significativos	2,10	1,90	2,30	2,27	1,93	2,61	2,26	1,97	2,55
17(37) «El error Tipo I se puede cometer cuando la hipótesis nula NO es rechazada»	3,00	2,70	3,30	2,72	2,35	3,09	3,91	3,62	4,20
22(42) «Es posible cometer a la vez error Tipo I y Tipo II en una prueba estadística»	2,80	2,50	3,10	3,37	3,00	3,74	3,53	3,22	3,84
29(49) «El cometer error Tipo II es bastante común en las investigaciones publicadas»	3,60	3,30	3,90	2,52	2,21	2,83	3,06	2,84	3,28

Notas:
– (*) valores estimados
– Mayor valor (media y límites) significa mayor grado de acierto
– Entre comillas: ítems considerados falsos (escala resp. fue invertida)

En general, los resultados con estos *ítems* han sido bastante buenos en las tres muestras, con niveles aceptables de acierto. Por ello podemos deducir que este tipo de error o falsa creencia de tomar los valores '*p*' como indicadores del tamaño de los efectos no tiene una presencia relevante entre los investigadores, en general.

La tabla 8 corresponde al octavo de los factores lógicos. Tres *ítems* que indagan el conocimiento sobre la relación entre el valor *'p'* y la importancia del resultado obtenido, los tres formulados como 'falsos'. Hemos de aclarar que el *ítem 18* nos ha suscitado serias dudas en su formulación y/o criterio de corrección. Así, desde nuestro punto de vista, que hemos comunicado a los autores de los otros dos estudios comparados, la redacción de dicho *ítem* debería ser «*los estudios con resultados estadísticamente no-significativos pueden ser aún muy importantes*», y cambiar el criterio de corrección de 'falso' a 'verdadero'. No obstante, para posibilitar la comparación, nos hemos ceñido a aplicar los mismos criterios seguidos en los otros dos estudios dejando el *ítem* en su redacción original, como aparece en la tabla, y puntuándolo como 'falso'.

Los datos bastante coincidentes en las tres muestras con respecto al *ítem* 18 tienen una lectura clara: no se consideran importantes los estudios con resultados estadísticamente no significativos, lo que en algunos casos puede ser verdad pero no necesariamente. La improbabilidad de un dato no es señal inequívoca de su importancia, entre otras razones porque un resultado no significativo en una nueva investigación con mayor tamaño de muestra puede alcanzar significación estadística. Se sabe, además, que los resultados nulos pueden ser en algunos casos interesantes por sí mismos o expresión fehaciente de la falta de potencia estadística. En resumen, en torno al concepto *p* hay mucha confusión porque le atribuimos, por lo general, más información de la que puede aportar.

Con respecto a los otros dos ítems, el nivel de acierto obtenido por la muestra española es superior a las estadounidenses, considerablemente en el ítem 6.

Tabla 6
Factor VI Percepciones sobre la influencia del tema o de la muestra

	Mittag y Thompson (2000)			Gordon (2001)			Nuestro (muestra de profesores Univ.)		
	Media (*)	**Límite inf.(*)**	**Límite sup.(*)**	**Media**	**Límite inf.**	**Límite sup.**	**Media**	**Límite inf.**	**Límite sup.**
10(30) Resultados estadísticamente significativos resultan más destacables cuando el tamaño de la muestra es pequeño	2,20	1,90	2,50	2,37	1,96	2,78	2,90	2,63	3,17
16(36) Toda hipótesis nula será finalmente rechazada con un tamaño de muestra determinado	3,00	2,70	3,30	3,15	2,75	3,55	2,55	2,29	2,81
25(45) Las pruebas de significación estadística indican, en parte, si el investigador ha trabajado o no con una muestra grande.	3,25	3,00	3,50	2,87	2,50	3,24	2,61	2,37	2,85

Notas:
– (*) valores estimados
– Mayor valor (media y límites) significa mayor grado de acierto
– Entre comillas: ítems considerados falsos (escala resp. fue invertida)

Tabla 7
Factor VII Percepción del valor p como medida del efecto

	Mittag y Thompson (2000)			Gordon (2001)			Nuestro (muestra de profesores Univ.)		
	Media (*)	**Límite inf.(*)**	**Límite sup.(*)**	**Media**	**Límite inf.**	**Límite sup.**	**Media**	**Límite inf.**	**Límite sup.**
11(31) «Valores de «p» pequeños ofrecen evidencia directa de que los efectos del tratamiento han sido grandes»	3,80	3,50	4,10	3,27	2,91	3,63	3,54	3,29	3,79
14(34) «Si una docena de investigadores estudiaron el mismo fenómeno usando la misma hipótesis nula, y ninguno de sus estudios arrojó resultados estadísticamente significativos, significa que los efectos investigados no son destacables ni importantes»	4,00	3,70	4,30	3,82	3,45	4,19	3,90	3,70	4,10
24(44) Los valores de «p» obtenidos en diferentes pruebas estadísticas no pueden ser directamente comparados, porque estos valores dependen de los tamaños de muestra utilizados en cada prueba	3,25	3,00	3,50	3,15	2,77	3,53	3,35	3,09	3,61

Notas:
– (*) valores estimados
– Mayor valor (media y límites) significa mayor grado de acierto
– Entre comillas: ítems considerados falsos (escala resp. fue invertida)

La siguiente tabla corresponde al noveno y último de los factores lógicos (véase tabla 9). Tres *ítems* que indagan sobre el conocimiento de la relación entre el valor *'p'* y la probabilidad de que los resultados se den en la población, dos formulados como 'falsos' y uno, el 15 como 'verdadero', basándose en una objeción que ha venido tomando fuerza en los últimos años de que las pruebas de significación estadística no evalúan la probabilidad de que los resultados obtenidos puedan ser replicados ni la probabilidad de que se den en la población (Cohen, 1994; Sohn, 1998; Thompson, 1996).

En este caso prácticamente las tres muestras coinciden en mostrar un suficiente nivel de acierto en el *ítem* 8, comparativamente mayor en la muestra española, y un bajo nivel de acierto en los dos restantes. Se pone de manifiesto la necesidad de clarificar el concepto de replicación, que no se debe confundir con el nivel de probabilidad.

Discusión y conclusiones

Del trabajo podemos derivar tres conclusiones generales y algunas recomendaciones tanto para la docencia como para la práctica investigadora.

La primera conclusión es la existencia de bastantes deficiencias de conceptualización relacionadas con el uso de la metodología estadística entre los investigadores en Psicología.

La segunda conclusión es que estas deficiencias, con algunos matices diferenciales, son básicamente coincidentes entre la muestra española y las norteamericanas estudiadas. En algunos factores, como se ha señalado oportunamente, la muestra española es superior en conocimiento correcto a las muestras norteamericanas y en otros casos, no.

Y la tercera conclusión es que de este estudio se derivan claves para conocer qué conceptos metodológico-estadísticos presentan

Tabla 8
Factor VIII Percepción del valor p como medida de la importancia del resultado

	Mittag y Thompson (2000)			**Gordon (2001)**			**Nuestro (muestra de profesores Univ.)**		
	Media (*)	**Límite inf.(*)**	**Límite sup.(*)**	**Media**	**Límite inf.**	**Límite sup.**	**Media**	**Límite inf.**	**Límite sup.**
6(26) «Un resultado con una p<0.05 indica que ese resultado es importante»	2,80	2,50	3,10	2,80	2,36	3,24	3,89	3,67	4,11
18(38) «Los estudios con resultados no-significativos pueden ser aún muy importantes»	1,40	1,10	1,70	1,45	1,08	1,82	1,99	1,81	2,17
27(47) «Los resultados improbables son generalmente los más importantes y destacables»	3,40	3,10	3,70	3,50	3,17	3,83	3,89	3,68	4,10

Notas:
– (*) valores estimados
– Mayor valor (media y límites) significa mayor grado de acierto
– Entre comillas: ítems considerados falsos (escala resp. fue invertida)

Tabla 9
Factor IX Percepción del valor p como evidencia de replicabilidad

	Mittag y Thompson (2000)			**Gordon (2001)**			**Nuestro (muestra de profesores Univ.)**		
	Media (*)	**Límite inf.(*)**	**Límite sup.(*)**	**Media**	**Límite inf.**	**Límite sup.**	**Media**	**Límite inf.**	**Límite sup.**
8(28) «Cuanto más pequeños son los valores de "p" más frecuentemente resultarán replicados dichos hallazgos en el futuro»	3,00	2,70	3,30	3,05	2,68	3,42	3,53	3,28	3,78
15(35) Los valores de «p» obtenidos en una investigación miden la probabilidad de que dichos resultados ocurran en la muestra, pero no la probabilidad de que ocurran en la población	3,00	2,70	3,30	2,82	2,41	3,23	2,34	2,09	2,59
21(41) «Las pruebas de significación estadística indican la probabilidad de que los resultados obtenidos con la muestra utilizada se den también en la población»	2,80	2,50	3,10	2,22	1,88	2,56	2,65	2,38	2,92

Notas:
– (*) valores estimados
– Mayor valor (media y límites) significa mayor grado de acierto
– Entre comillas: ítems considerados falsos (escala resp. fue invertida)

mayores deficiencias y cuáles menos en la muestra española, lo que debería servir para mejorar la docencia en este campo y en la incorporación de algunos componentes de la corriente crítica en los programas docentes de las universidades, máxime ahora que España se encuentra en el inicio del proceso de homologación de los estudios superiores con el resto de Estados de la Unión Europea (*The European Higher Education Area*).

De esta tercera conclusión podemos derivar algunas recomendaciones (para antes de publicar un trabajo):

I) Con respecto al uso de las Pruebas de Significación Estadística:

– La conveniencia de utilizar siempre la frase «estadísticamente significativo» en lugar de sólo «significativo», a efectos de evitar confusión entre la estricta significación estadística y la significación social (sustantiva, práctica, clínica) de los datos (Carver, 1993; Gliner, Morgan, Leech y Harmon, 2001).
– Ofrecer siempre alguna estimación del tamaño del efecto cuando se informe de un valor *'p'*. La postura oficial adoptada por la APA afirma que las pruebas de significación no se deben abandonar, pero deberán complementarse con información sobre el tamaño del efecto (Frías, Pascual y García, 2000; Wilkinson y The APA Task Force on Statistical Inference, 1999, p. 599). Valga la recomendación general de Thompson (1998) que en los estudios en los que se comprueben efectos siempre es mejor aportar «alguna» información sobre el tamaño de los efectos que «*ninguna*».
– Atajar la confusión bastante extendida entre los investigadores de que el no-rechazo de la hipótesis nula implica la veracidad de la hipótesis nula. Finch y sus colaboradores, revisando los artículos publicados en el *Journal of Applied Psychology* durante los últimos sesenta años, encontraron que en el 38% de ellos los resultados estadísticamente no-significativos eran interpretados como que la hipótesis nula se consideraba verdadera (Finch et al., 2001).
– Emplear hipótesis nulas diferentes de '0' (Cohen, 1994; Frías, Pascual y García, 2002; Frías, Pascual y Monterde-Bort, 2005; Thompson, 1999a).

II) Con respecto al Modelo Lineal General:

– Desterrar la falsa creencia, histórica en la Psicología, de que la regresión/correlación es una técnica solamente adecuada para la investigación «no-experimental» y que el análisis de varianza caracteriza la investigación «experimental». Siendo equivalentes, la regresión/correlación tiene una ventaja añadida al aportar medidas del tamaño de los efectos: una importante implicación del M.L.G. es que r^2 y sus análogos (η^2, ω^2…) se pueden utilizar como una estimación del tamaño del efecto *en todos los análisis* (Mittag y Thompson, 2000, p.15; Thompson, 2000, p. 2).

III) Con respecto a los procedimientos paso-a-paso:

– No interpretar el orden de selección de las variables predictoras como indicador de la importancia de éstas, sino utilizar como alternativa los *coeficientes estandarizados (Beta)* o los *coeficientes de estructura* (Thompson y Borello, 1985). Los métodos paso-a-paso han recibido numerosas críticas en su capacidad para identificar el mejor conjunto de variables predictoras (Carver, 1978; Snyder, 1991; Thompson, 1994a, 1994b, 1998).

IV) Con respecto a la fiabilidad de los instrumentos de medida:

– Sería conveniente establecer la distinción entre el concepto de fiabilidad y el coeficiente o índice obtenido, es decir, entre lo que queremos medir y lo que logramos medir. Los índices de fiabilidad obtenidos lo son de los datos ofrecidos por un test para una población particular de examinados (Feldt y Brennan, 1989; Mittag y Thompson, 2000; Wilkinson and The APA Task Force on Statistical Inference, 1999) y en un momento temporal determinado. Es importante, por tanto, que los investigadores ofrezcan para cada estudio datos sobre la fiabilidad de las puntuaciones obtenidas, incluso cuando el objetivo de su estudio no sea psicométrico.

V) Con respecto a la replicabilidad:

– Se debería de ofrecer alguna garantía de que los resultados presentados son replicables (Thompson, 1998, p. 16, y 1999b, p. 173). El valor de *p* no es un predictor concluyente de la replicabilidad de los datos (Pascual, García y Frías, 2000). Sin embargo, la replicabilidad de cualquier hallazgo es esencial en la ciencia. Es necesario difundir el conocimiento de estrategias de replicación interna (*jackknife* y *bootstrap*, entre otras) y de replicación directa y sistemática para evitar el probable sesgo en el uso habitual de muestras de conveniencia.

Agradecimientos

Este trabajo forma parte de los resultados de un proyecto subvencionado por el Ministerio de Ciencia y Tecnología bajo el título «*Revisión crítica del diseño estadístico en Psicología: análisis de los programas docentes y protocolos editoriales de revisión*». (Ref: SEJ2004-08304).

También agradecemos al Colegio Oficial de Psicólogos de la Comunidad Valenciana su colaboración en la obtención de la muestra.

Referencias

American Psychological Association (2001). *Publication manual of the American Psychological Association (5th ed.)*. Washington, DC: APA.

Bakan, D. (1966). The test of significance in psychological research. *Psychological Bulletin, 66*, 423-437.

Borges, A., San Luis, C., Sánchez, J.A., y Cañadas, I. (2001). El juicio contra la hipótesis nula: muchos testigos y una sentencia virtuosa. *Psicothema, 13*, 173-178.

Bracey, G.W. (1988). Tips for readers of research. *Phi Delta Kappan, 70*, 257-258.

Carver, R.P. (1978). The case against statistical significance testing. *Harvard Educational Review, 48*, 378-399.

Carver, R.P. (1993). The case against statistical significance testing revisited. *Journal of Experimental Education, 61,* 287-292.

Cohen, J. (1968). Multiple regression as a general data-analytic system. *Psychological Bulletin, 70*, 426-443.

Cohen, J. (1990). Things I have learned (so far). *American Psychologist, 45*(12), 1304-1012.

Cohen, J. (1994). The earth is round (p<.05). *American Psychologist, 49*(12), 997-1003.

Cohen, J., West, S.G., Cohen, P., y Aiken, L. (2003). *Applied multiple regression/correlation analysis for the behavioral sciences* (3rd ed). Mahwah, N.J.: LEA, Inc.

Cox, D.R. (1977). The role of significant tests. *Scandinavian Journal of Statistics, 4*, 49-70.

Falk, R. (1998). In criticism of the null hypothesis statistical test. *American Psychologist, 53,* 798-799.

Feldt, L.S., y Brennan, R.L. (1989). Reliability. En R.L. Linn (ed.): *Educational measurement* (3rd ed., pp. 105-146). Washington, DC: American Council on Education.

Finch, S., Cumming, G., y Thomason, N. (2001). Reporting of statistical inference in the Journal of Applied Psychology: Little evidence of reform. *Educational and Psychological Measurement, 61,* 181-210.

Frías, M.D., Pascual, J., y García, J.F. (2000). Tamaño del efecto del tratamiento y significación estadística. *Psicothema, 12* (suplem. 2), 236-240.

Frías, M.D., Pascual, J., y García, J.F. (2002). La hipótesis nula y la significación práctica. *Metodología de las Ciencias del Comportamiento, 4* (especial), 181-185.

Frías, M.D., Pascual, J., y Monterde-Bort, H. (2005). Reform of statistical practice in psychology. *9th European Congress of Psychology,* EFPA. Granada (Spain), July-2005.

Gliner, J.A., Morgan, G.A., Leech, N.L., y Harmon, R.J. (2001). Problems with null hypothesis significance tests. *Journal of the American Academy of Child and Adolescent Psychiatry, 40,* 250-252.

Gordon, H.R.D. (2001). American Vocational Education Research Association members' perceptions of statistical significance tests and other statistical controversies. *Journal of Vocational Educational Research, 26*(2), 1-18.

Harding, S. (1976). *Can Theories Be Refuted? Essays on the Duhem-Quine thesis*. Dordrecht, the Netherlands: D. Reidel.

Huberty, C.J. (1993) Historical origins of statistical testing practices: The treatment of Fisher versus Neyman-Pearson views in textbooks. *Journal of Experimental Education, 671*, 317-333.

Kerlinger, F.N. (1986). *Foundations of behavioral research.* New York, NY: Holt, Rinehart y Winston.

Krueger, J. (2001). Null hypothesis significance testing: On the survival of a flawed method. *American Psychologist, 56,* 16-26.

Mittag, K.C. (1999). The psychometrics group instrument: Attitudes about contemporary statistical controversies. Unpublished instrument. The University of Texas at San Antonio.

Mittag, K.C., y Thompson, B. (2000). A national survey of AERA members' perceptions of statistical significance tests and other statistical issues. *Educational Researcher, 29*(4), 14-20.

Monterde-Bort, H., Pascual, J., y Frías, M.D. (2005). Incomprensión de los conceptos metodológicos y estadísticos: La encuesta «USABE». *IX Congreso de Metodología de las Ciencias Sociales y de la Salud.* Granada (Spain), septiembre 2005.

Moore, G.E. (1992). The significance of research in Vocational Education: The 1992 AVERA presidential address. *Journal of Vocational Education Research, 17*(4), 1-4.

Morgan, P.L. (2003). Null hypothesis significance testing: Philosophical and practical considerations of a statistical controversy. *Exceptionality, 11*(4), 209-221.

Nickerson, R.S. (2000). Null hypothesis significance testing: A review of old and continuing controversy. *Psychological Methods, 5*(2), 241-301.

Pascual, J., García, J.F., y Frías, M.D. (2000). Significación estadística, importancia del efecto y replicabilidad de los datos. *Psicothema, 12* (suplem. 2), 408-412.

Popper, K. (1963). *Conjectures and refutations*. New York, NY: Harper & Row.

Rozeboom, W.W. (1997). Good science is adductive, not hypothetic-deductive. En Harlow, L.L., Mulaik, S.A., y Steiger, J.H. (eds.): *What if there were no significance tests?* Mahwah, N.J.: LEA, Inc.

Schmidt, F.L., y Hunter, J.E. (2002). Are there benefits from NHST? *American Psychologist, 57,* 65-66.

Snyder, P. (1991). Three Reasons Why Stepwise Regression Methods Should Not Be Used by Researchers. In B. Thompson (ed.): *Advances in Educational Research: Substantive Findings, Methodological Developments* (vol. 1, pp. 99-105). Greenwich, CT: JAI Press.

Sohn, D. (1998). Statistical significance and replicability. *Theory & Psychology, 8,* 291-311.

Thompson, B. (1994a). The concept of statistical significance testing (An ERIC/AE Clearinghouse Digest EDO-TM-94-1). *Measurement Update, 4*(1), 5-6.

Thompson, B. (1994b). Guidelines for authors. *Educational and Psychological Measurement, 54,* 837-847.

Thompson, B. (1996). AERA editorial policies regarding statistical significance testing: Three suggested reforms. *Educational Researcher, 25*(2), 26-30.

Thompson, B. (1998). *Five methodology errors in educational research: The pantheon of statistical significance and other faux pas.* Annual meeting of the American Educational Research Association, San Diego, April 15, 1998. Documento disponible en www.coe.tamu.edu/~btthompson/aeraaddr.htm

Thompson, B. (1999a). *Common methodology mistakes in educational research, revisited*. Annual meeting of the American Educational Research Association, Montreal, April 22, 1999. Documento disponible en www.coe.tamu.edu/~btthompson/aeraad99.htm.

Thompson, B. (1999b). If statistical significance tests are broken/misused, what practices should supplement or replace them? *Theory & Psychology, 9,* 167-183.

Thompson, B. (2000). *A suggested revision to the forthcoming 5th edition of the APA publication manual.* Documento disponible en http://www.coe.tamu.edu/~bthompson/apaeffec.htm.

Thompson, B., y Borello, G.M. (1985). The importance of structure coefficients in regression research. *Educational and Psychological Measurement, 45,* 203-209.

Wilkinson, L., y The APA task force on statistical inference (1999). Statistical methods in psychology journals guidelines and explanations. *American Psychologist, 54*(8), 594-604.