Psicothema
2000. Vol. 12, Supl. nº 2, pp. 408-412

ISSN 0214 - 9915 CODEN PSOTEG

Significación estadística, importancia del efecto y replicabilidad de los datos

Juan Pascual Llobell, José Fernando García Pérez y María Dolores Frías Navarro
Universidad de Valencia

Se analiza la relación entre los conceptos de significación estadística (nivel de probabilidad, *p*) y de replicabilidad. El nivel de significación estadística (p. e. de 0.01) indica la probabilidad de los datos bajo el supuesto de la hipótesis nula, pero eso no quiere decir que una replicación posterior tendrá la probabilidad complementaría (en este caso, 0.99) de ser significativa. Entendida correctamente la replicabilidad tiene que ver exclusivamente con la fiabilidad y consistencia de los datos, y la única forma de comprobarla es mediante sucesivos contrastes empíricos.

Statistical significance and replicability of the data. This paper analyses the relationship between the concepts of statistical significance (level of probability, p) and replicability. The level of statistical significance (for example, p = 0.01) indicates the probability of the data under the null hypothesis assumption, however, this does not mean that in a later replication the probability to obtain significant differences will be the complementary, 0.99. If correctly understood, replicability is exclusively related to the reliability and consistency of the data. The only way to evaluate reliability is through repeated empirical tests.

Desde hace muchos años la práctica de la experimentación en psicología, especialmente en su modalidad de investigación de laboratorio, está asociada al contraste y comprobación de hipótesis estadísticas *(null hypothesis significance testing*), sobre todo desde la introducción del análisis de la variancia por Fisher (1925), dado que el análisis estadístico propuesto aportó un criterio de decisión simple y suficiente, el valor *p* de probabilidad, como referente último del rechazo/aceptación de las hipótesis teóricas sometidas al proceso de verificación empírica.

Con el paso del tiempo, la asociación establecida entre experimentación y análisis estadístico fomentó, quizá en exceso, una práxis profesional que derivó en norma de obligado cumplimiento y estándar de publicación, a pesar de que desde una plataforma más teórica y analítica dicha práctica fue sistemáticamente cuestionada por insuficiente unas veces, por errónea en su interpretación otras o simplemente por considerársela responsable, junto con otros factores, de conducir a la psicología por unas rutinas de trabajo que habiendo potenciado la productividad científica apenas han logrado lo que es propio de toda ciencia: la «acumulación de conocimiento». (Entre otros, véanse las monografías de Chow (1996), Harlow, Mulaik y Steiger (1997) y Schmidt (1996)).

En este artículo revisamos algunas de las críticas que se han formulado acerca del uso/abuso del valor *p* (probabilidad del estadístico de contraste, sea *F*, *t*, …), entre ellas, la de interpretarlo como sinónimo de la replicabilidad (consistencia) de los datos o la de considerarlo por el contrario como un indicador de escasa utilidad informativa por cuanto la hipótesis nula a comprobar es siempre falsa y, en consecuencia, su rechazo no aporta nada que no sea previamente sabido. De las dos anteriores, la primera afirmación quizá peca por exceso y la segunda probablemente por defecto. Las tesis que nos atrevemos a proponer en este artículo se oponen diametralmente a las anteriores porque consideramos que: *a*) *p* no (siempre ni necesariamente) es un indicador del valor de replicabilidad de los datos y *b*) el proceso de comprobación de la hipótesis nula sí es informativo porque la hipótesis nula, al menos en algunos casos, se la puede considerar como explicación plausible.

Antes de argumentar a favor de ambas tesis convendría delimitar, a efectos aclarativos, el contexto teórico de argumentación. Al comprobar hipótesis estadísticas acerca de algún parámetro (por ejemplo, la diferencia entre dos medias muestrales, o lo que es lo mismo, si dos muestras pertenecen o no a la misma población, -recuérdese que «una hipótesis estadística es siempre una afirmación sobre la población, no sobre la muestra», Hayes, 1963, pág. 248, se parte siempre de la hipótesis de nulidad: en este caso, la no-existencia de diferencias. Esta hipótesis de nulidad cumple el *status* epistemológico de definir la función de probabilidad del estadístico de referencia (*F*, *t*, …); es decir, la hipótesis nula delimita la existencia de un mundo supuesto (posible) en el que se cumplen las características de la distribución del test estadístico elegido de manera tal que, conocida su distribución, es posible determinar con exactitud la probabilidad asociada al mismo. En definitiva, el valor *p* define la probabilidad de los datos bajo el supuesto de verdad de la hipótesis nula, hipótesis que en la mayoria de los casos es la única que se puede someter a comprobación, «*because we can never know the true population parameter when* H_1 *is true*» (Hagen, 1997, pág. 17).

Correspondencia: Juan Pascual Llobell
Facultad de Psicología
Universidad de Valencia
46010 Valencia (Spain)
E-mail: Juan.Pascual@uv.es

Acerca de las relaciones entre *p* y replicabilidad han coexistido, al menos de hecho, posturas encontradas. Hace años Bakan (1966) afirmó que el valor *p* no era una medida adecuada de la fiabilidad (replicabilidad) de los resultados obtenidos. En los mismos planteamientos se pronunció más ampliamente Lykken (1968) y últimamente, Gigerenzer (1993) apeló a la por él llamada «*falacia de la replicación*», consistente en creer que cuanto mayor es el nivel de significación estadística mayor es la probabilidad de que los resultados sean replicables en una futura investigación o experimento.

A pesar de ello se ha podido constatar sociológicamente, al menos si nos atenemos a los resultados de una encuesta entre expertos descrita por Oakes (1986), que el 60% de los investigadores consideran la afirmación siguiente como cierta: «*supuesto que se haya obtenido un valor de p = 0.01, al repetir el experimento un gran número de veces, obtendremos resultados significativos en el 99% de los casos» (p. 173).*

Los encuestados a todas luces confundían el valor *p* con la potencia de la prueba estadística: Suponiendo, según el ejemplo anterior, que de la aplicación de la prueba estadística obtenemos un valor $t = \sqrt{N}\ 2.7$, con g.l. = 38; y sabiendo que el tamaño del efecto (*d*) es igual a $2t/\sqrt{N}$, por tanto igual a 0.85, se deriva una potencia estimada de 0.43. Esto es, la probabilidad de encontrar el mismo resultado al repetir el experimento es del 43% y en ningún caso del 99% que le atribuían los encuestados.

Recientemente Greenwald, González, Harris, y Guthrie, (1996) al intentar encontrar razones que justifiquen el uso continuado pese a todo, del contraste de hipótesis concluían que dicho valor sí proporciona una indicación válida de la replicabilidad de la decisión tomada en contra de la hipotesis nula:

> «*Although we agree with most critics' catalogs of NHT's flaws, this article also takes the unusual stance of identifying virtues that may explain why NHT continues to be so extensively used. These virtues include providing results in the form of dichotomous (yes/no) hypothesis evaluation and providing and index (p value) that has a justifiable mapping onto* ***confidence in repeatability of a null hypothesis rejection***» (pág. 179. El remarcado es nuestro).

Los autores entienden que replicar consiste en *generar un nuevo rechazo de la hipótesis nula* manteniendo constante las condiciones de observación. Operativamente, la definen como:

$$1-\beta = 1-P\left(\frac{t_{crít}-t_1}{\sqrt{1+\frac{t_{crít}^2}{2\times gl}}}\right) \qquad [1]$$

Donde $t_{crít}$ es el valor de *t* necesario para rechazar la hipótesis nula con *gl* grados de libertad, *P* es la probabilidad acumulada de la distribución normal, y t_1 el valor de *t* obtenido en el primer estudio. Según esto, unos resultados con un valor *p* de 0.005 deberían ser más replicables que otros con valor *p* = 0.01. En general, concluyen los autores, cuando la probabilidad asociada a los datos es de 0.05 la probabilidad de replicación está en torno al 50% (si se aplica la fórmula anterior se obtendrá exactamente este valor) y si es de 0.005 la probabilidad de replicación será de 0.80, aproximadamente el valor convencional de potencia deseada y conveniente según el criterio autorizado de Cohen (1977).

El razonamiento de los autores parece impecable pero también, añadimos nosotros, es incompleto, pues parten de un concepto de «replicación» insuficiente a todas luces*: los autores sólo contemplan el caso en el que la hipótesis nula es falsa,* es decir cuando de hecho existe un efecto experimental. Dado que existe, la tarea del cientifico debe consistir en detectarlo y estimar sus tamaños. Según algunos autores esto es así y no puede ser de otra manera porque la hipótesis nula *siempre es falsa*. Así piensan entre otros Meehl (1967, 1990) y Cohen (1990, 1994). Pero si así fuera, se podría gualmente concluir que el procedimiento de contraste de hipótesis es del todo improcedente o redundante, ¿para qué comprobar lo que claramente ya se sabe que es y existe? Lo sensato en buena lógica sería abandonar tal estrategia de investigación.

Es verdad que la hipótesis nula puede ser falsa en ciertos casos pero eso no quiere decir que necesariamente tenga que ser así. Dos muestras obtenidas a partir de la misma población, eso es lo que debe suponer la hipótesis nula, siempre podrán diferir entre sí; si la variable en cuestión es medida con «precisión infinita» los grupos de sujetos muestrales siempre diferirán *algo* entre sí. Los grupos de sujetos sólo podrían llegar a ser iguales en el caso extremo de que el tamaño muestral fuera igual al tamaño poblacional. Precisemos por esto la hipótesis de nulidad no puede entenderse como una hipótesis acerca de la existencia de diferencias entre dos grupos o condiciones, porque la hipótesis nula no tiene que ver con las diferencias muestrales, que de hecho casi siempre existirán, sino que supuestas esas diferencias y a pesar de ellas, la hipótesis nula se pregunta si ambos grupos o muestras pertenecen a la misma población y con qué probabilidad.

Por tanto, afirmar que la hipótesis de nulidad es siempre falsa sin más, no puede ser verdad. Siempre podremos demostrar la falsedad de la hipótesis de nulidad cuando sea falsa, eso sí es cierto: poder demostrar la falsedad de algo si efectivamente lo es, no es lo mismo que suponer de partida que ese algo siempre es falso. En consecuencia, la hipótesis nula considerada a priori puede ser tanto verdadera como falsa. Supongamos por un momento que puede ser verdadera; si así fuera ¿el valor *p de los datos bajo ese supuesto también sería un indicador válido de la replicabilidad de los datos? Esa es la pregunta relevante que hay que hacerse para responder de manera definitiva sobre la relación entre p y replicabilidad.*

Simulemos un ejemplo ficticio en el que se pueda presumir que la hipótesis nula es plausible (verdadera). Supongamos un alumno *completamente* ignorante de una materia determinada de examen (dadas las condiciones acuales de docencia y del sistema de exámenes imperante no es tan descabellado suponerlo); operativamente podemos conseguir que esto sea absolutamente cierto haciendo que responda a un examen con cuatro alternativas cerradas sin tener conocimiento de las preguntas. La hipótesis de partida es que su nivel de conocimiento es nulo, en consecuencia, la hipótesis nula será cierta, luego $p(H_0) = 1$. En la *Tabla 1* representamos la función de probabilidad, la de distribución y el valor *p* de cada suceso.

Para determinar el valor *p* a partir del cual decidimos rechazar la hipótesis de nulidad fijamos un limite α de 0.10. (Como el espacio muestral de respuesta varía entre 0-20 aciertos, el valor más aproximado a 0.10 corresponde a 8 aciertos).

Tabla 1
Función de probabilidad, función de distribución y valor p de la distribución binomial (20, 0.25)

n	*x*	*f(x)*	*F(x)*	*p*
20	0	0.003171212	0.003171212	1.000000000
20	1	0.021141413	0.024312625	0.996828788
20	2	0.066947808	0.091260432	0.975687375
20	3	0.133895615	0.225156048	0.908739568
20	4	0.189685455	0.414841503	0.774843952
20	5	0.202331152	0.617172654	0.585158497
20	6	0.168609293	0.785781948	0.382827346
20	7	0.112406195	0.898188143	0.214218052
20	8	0.060886689	0.959074832	**0.101811857**
20	9	0.027060751	0.986135583	0.040925168
20	10	0.009922275	0.996057858	0.013864417
20	11	0.003006750	0.999064608	0.003942142
20	12	0.000751688	0.999816296	0.000935392
20	13	0.000154192	0.999970488	0.000183704
20	14	0.000025699	0.999996187	0.000029512
20	15	0.000003426	0.999999613	0.000003813
20	16	0.000000357	0.999999970	0.000000387
20	17	0.000000028	0.999999998	0.000000030
20	18	0.000000002	1.000000000	0.000000002
20	19	0.000000000	1.000000000	0.000000000
20	20	0.000000000	1.000000000	0.000000000

Si la hipótesis nula es cierta, la distribución teórica se ajustará perfectamente a la que hemos elaborado (véase *Tabla 1*). La probabilidad de que un alumno no acierte ningún elemento $f(x = 0)$ será de 0.003, que acierte solamente 1 será (0.021), que acierte 2 es $f(x = 2) = 0.067$, y así sucesivamente. Si para aprobar fuera necesario obtener una puntuación exacta de 6, se puede afirmar que la probabilidad de aprobar sin saber nada sería de 0.169. Pero si para aprobar se precisara obtener la puntuación 6 u otra cantidad mayor, como suele ocurrir en los exámenes, la probabilidad de que $F(x \geq 6) = f(x = 6) + f(x = 7) + \ldots + f(x = 20)$, por tanto, $F(x \geq 6) = 0.382$; cantidad que corresponde con el valor de *p* en este punto de la distribución, $p(x \geq 6) = 1 - F(x < 6) = 0.382$.

De acuerdo con esta lógica, simulemos ahora que cinco millones de alumnos responden independientemente al mismo examen dos veces consecutivas,(esto es, una vez y su réplica) sin conocer cuáles son las preguntas del mismo. La simulación se ha realizado con la función *RV.BINOM(N, p)* del programa SPSS, fijando los parámetros en 20 y 0.25, respectivamente. Los resultados se muestran en la *Tabla 2*. Las filas de la matriz definen la primera vez y las columnas las réplicas.

Según Greenwald y colaboradores (1996) cuando el valor *p* es de 0.005, la probabilidad de replicar el mismo resultado es del 80%, y a partir de este valor de *p*, la proporción de replicas irá en aumento. Para comprobar si se cumple esta predicción en los datos simulados, se resume en la *Tabla 3* la distribución teórica y las probabilidades asociadas a cada suceso. En negrita hemos marcado los casos en los que se rechaza la hipótesis de nulidad con el nivel α fijado previamente. En la parte de la tabla correspondiente a las réplicas, las dos primeras columnas se corresponden con el no-rechazo de la hipótesis nula y las dos segundas con el rechazo. Se puede comprobar que el rechazo de la hipótesis nula en la réplica es independiente del rechazo en el primer experimento. Así, si la probabilidad del primer rechazo fue de 0.10 (primera columna, fila 8), la probabilidad de rechazo en la réplica fue del 10.22%, pero si la probabilidad de rechazo en el primer experimento fue de 0.0009, el rechazo en la replica igualmente se mantiene en torno a 10.15% (En todos los casos el porcentaje de rechazos coincide con el alpha).

Luego el valor *p* no es un predictor concluyente de la replicabilidad de los datos. Asumir lo contrario podría llevarnos a cometer algún error grave de interpretación: *al encontrarse el investigador con un valor p muy bajo en un primer experimento podría, dado que cree que dicho valor bajo es representativo de una alta replicabilidad y consistencia del efecto, concluir que el resultado es concluyente cuando en realidad pudo haber cometido simplemente un error de Tipo I. Que el error Tipo I sea pequeño, supongamos del .001 no quiere decir que no haya sido cometido precisamente en este experimento.*

Tabla 2
Tabla de contingencia de 5 millones de réplicas de un estudio cuando la hipótesis nula es cierta, de una distribución binomial (20 ensayos y probabilidad de acierto .25)

	0	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**F**	**%**
0	34	343	1096	2202	3034	3193	2686	1734	965	408	167	62	9	1	1			15935	.3
1	336	2287	7015	14102	19976	21528	17890	11923	6473	2844	1090	310	62	12	3	1		105852	2.1
2	1054	7037	22343	44895	63724	67715	56528	37640	20349	9177	3368	998	235	66	6	3		335138	6.7
3	2040	14099	44770	89563	127358	135299	112769	75777	40905	18138	6620	1986	520	105	15	3		669967	13.4
4	2923	19862	63391	126891	180075	191763	159625	107001	57551	25599	9566	2869	729	145	29	5		948024	19.0
5	3231	21284	67930	135514	192202	205521	171107	113896	61302	27280	10053	3102	739	156	21	1		1013339	20.3
6	2739	17725	56044	112823	159616	169831	141825	94568	51465	22889	8231	2436	622	159	25	10		841008	16.8
7	1805	11729	37807	75159	106727	113671	94594	63027	34172	15158	5585	1776	431	93	12	1		561747	11.2
8	973	6485	20675	40679	57910	61628	51057	34059	18772	8142	3007	913	242	46	8		1	304597	6.1
9	434	2876	9124	18164	25788	27344	22689	15187	8186	3598	1345	391	98	13	1			135238	2.7
10	164	1063	3285	6686	9337	10044	8365	5536	2953	1334	476	165	34	13	1	1		49457	1.0
11	51	313	1022	2137	2935	3004	2463	1684	888	410	136	47	18	4				15112	.3
12	15	88	219	493	696	801	588	421	218	115	32	7	3					3696	.1
13	3	14	46	123	139	160	123	72	47	20	8	2						757	.0
14	1	2	7	18	19	25	23	11	10	4								120	.0
15		1		2	1	6		1										11	.0
16					1	1												2	.0
F	15803	105208	334774	669451	949538	1011534	842332	562537	304256	135116	49684	15064	3742	813	122	25	1	5000000	
%	.3	2.1	6.7	13.4	19.0	20.2	16.8	11.3	6.1	2.7	1.0	.3	.1	.0	.0	.0	.0		100.0

Tabla 3
Proporción de réplicas de la distribución binomial (20, .25) donde se rechaza la hipótesis nula

Binomial (20, .25)				*1º*	*Réplica*			
					p > 0.10		*p £ 0.10*	
Aciertos	**p**	**f(X)**	**%**	**Total**	**f**	**%**	**f**	**%**
0	1.000000	0.003171	0.003187	15935	14322	89.878	1613	10.12
1	0.996829	0.021141	0.021170	105852	95057	89.802	10795	10.20
2	0.975687	0.066948	0.067028	335138	300936	89.795	34202	10.21
3	0.908740	0.133896	0.133993	669967	601675	89.807	68292	10.19
4	0.774844	0.189685	0.189605	948024	851531	89.822	96493	10.18
5	0.585158	0.202331	0.202668	1013339	910685	89.870	102654	10.13
6	0.382827	0.168609	0.168202	841008	755171	89.794	85837	10.21
7	0.214218	0.112406	0.112349	561747	504519	89.812	57228	10.19
8	***0.101812***	***0.060887***	***0.060919***	***304597***	***273466***	***89.780***	***31131***	***10.22***
9	***0.040925***	***0.027061***	***0.027048***	***135238***	***121606***	***89.920***	***13632***	***10.08***
10	***0.013864***	***0.009922***	***0.009891***	***49457***	***44480***	***89.937***	***4977***	***10.06***
11	***0.003942***	***0.003007***	***0.003022***	***15112***	***13609***	***90.054***	***1503***	***9.95***
12	***0.000935***	***0.000752***	***0.000739***	***3696***	***3321***	***89.854***	***375***	***10.15***
13	***0.000184***	***0.000154***	***0.000151***	***757***	***680***	***89.828***	***77***	***10.17***
14	***0.000030***	***0.000026***	***0.000024***	***120***	***106***	***88.333***	***14***	***11.67***
15	***0.000004***	***0.000003***	***0.000002***	***11***	***11***	***100.000***	***0***	***0.00***
16	***0.000000***	***0.000000***	***0.000000***	***2***	***2***	***100.000***	***0***	***0.00***
				5000000	**4491177**	**89.824**	**508823**	**10.18**

Predecir la replicabilidad a partir del valor de *p* únicamente funciona si suponemos (sabemos) que la hipótesis nula es falsa, pero como hemos afirmado anteriormente puede haber razones de índole teórica y también de índole estadística para pensar que no siempre es así. Más aún, nunca podremos estar seguros de no haber cometido el error de Tipo I en un momento dado; por tanto preguntarse si es verdadera o falsa la hipótesis nula es improcedente en este enfoque. De hecho, nunca lo sabremos a priori con seguridad. Para defender el contraste de hipótesis estadísticas no hay que situarse sólo en el caso más favorable (que sería la postura de estos autores) sino que es obligado contemplar además la posibilidad de que la hipótesis nula sea cierta. Obtener un valor *p* de 0.01 en un experimento dado puede ocurrir tanto cuando la hipótesis nula es cierta como cuando es falsa.

Una hipótesis nula, sea cierta o falsa, puede producir cualquier valor de *p*. En lo único que varían ambas es en la distribución de probabilidades. Por tanto, no puede recomendarse a los investigadores que confíen en que el valor de *p* sea un indicador de la replicabilidad a no ser que sepan con seguridad que la hipótesis de nulidad es falsa, en cuyo caso sería improcedente pasar a la comprobación de la hipótesis nula que ya de antemano se sabe que es falsa. En estos casos parece más oportuno estimar directamente otros parámetros de interés, por ejemplo, el tamaño del efecto, antes que empeñarse en comprobar la significación estadística.

Conclusiones

Concluimos, pues, que es engañoso afirmar sin más que la replicabilidad y el valor de *p* son lo mismo. Es cierto que en determinadas condiciones el valor de *p* puede funcionar como indicador de la mayor o menor replicabilidad de los datos, detectándose entre ambos valores una función monotónica demostrada por Greenwald y colaboradores (1996).

La relación monotónica creciente entre replicabilidad y *p* no se mantiene si suponemos que la hipótesis de nulidad es verdadera, según hemos razonado anteriormente. Un reflexión teórica más fina nos predispone a afirmar que la fiabilidad de un efecto es algo probabilísticamente desconocido y no existe método más objetivo de saber si un fenómeno, por ejemplo, una diferencia entre dos medias, es fiable que la replicación empírica del mismo. Los efectos fiables serán repetibles en posteriores observaciones independientes, mientras que los efectos aleatorios no lo serán (Hammond, 1996).

La replicación de cualquier hallazgo de investigación es esencial en la ciencia. Por ello es conveniente recordar las palabras de Thompson (1996): «*If science is the business of discovering replicable effects, because statistical significance test do not evaluate result replicability, then researchers should use and report some strategies that do evaluate the replicability of their results*» (pág. 29), entre las cuales, están las llamadas estrategias de «replicación externa» (realización de nuevos experimentos) y estrategias de «replicación interna» (los procedimientos *jackknife* y *bootstrap*). Desgraciadamente estos procedimientos no son, hoy por hoy, de uso común.

Referencias

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

Chow, S. L. (1996). *Statistical significance, rationale, validity and utility*. London: Sage.

Cohen, J. (1977). *Statistical power analysis for the behavioral sciences (ed. rev.)*. New York: Academic Press.

Cohen, J. (1990). Things I have learned (so far). *American Psychologist, 45*, 1304-1312.

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

Fisher, R. A (1925). *Statistical methods for research workers*. London: Oliver & Boyd.

Frick, R. (1995). On accepting the null hypothesis. *Memory & Cognition*, *23*, 132-138.

Frick, R. (1996). The appropriate use of null hypothesis testing. *Psychological Methods*, *1*, 379-390.

Gigerenzer, G. (1993). The superego, the ego, and the id in statistical reasoning. En G. Kereng y C. Lewis (eds.), *A handbook of data analysis in behavioral sciences: methodological issues* (pp. 311-339). Hillsdale, NJ: Lawrence Erlbaum Associates.

Greenwald, A. G., González, R., Harris, R. J., y Guthrie, D. (1996). Effect sizes and p values: what should be reported and what should be replicated? *Psychophysiology*, *33*, 175-183.

Hagen, R. L. (1997) In praise of the null hypothesis statistical test, *American Psychologist*, *52*, 15-24.

Hammond, G. (1996). The objections to null hypothesis testing as a means of analyzing psychological data. *Australian Journal of Psychology*, *2*, 104-106.

Harlow, L. L., Mulaik, S. A., y Steiger, J. H. (1997). *What if there were non significance tests?* London: Lawrence Erlbaum Associates.

Hayes, W. L. (1963). *Statistics for psychologists*. New York, N.Y.: Holt, Rinehart & Winston.

Lykken, D. T. (1968). Statistical significance in psychological research. *Psychological Bulletin*, *70*, 151-159.

Meehl, P. E. (1967). Theory-testing in psychology and phisics: A methodological paradox, *Philosophy of Science*, *34*, 103-115.

Meehl, P. E. (1990). Why summaries of research on psychological theories are often uninterpretable. *Psychological Reports*, *66*, 195-244.

Oakes, M. (1986). *Statistical inference: a commentary for social and behavioral sciences*. Chichester: John Wiley & Sons.

Schmidt, F. L. (1996). Statistical significance testing and cumulative knowledge in psychology: Implications for training of researchers. *Psychological Methods*, *1*, 115-129.

Thompson, B. (1966). AERA editorial policies regarding statistical significance testing: Three suggested reforms. *Educational Researcher*, *25*, 26-30.