p-hacking, hipótesis múltiples y replicabilidad;

Francesco Bogliacino

Back to Epistemology

- Quisiéramos que nuestras regresiones sean capaces de decir la verdad;
- Piensen en un teorema, o es cierto o no lo es. La verdad de una implicación lógica no depende del observador;
- En algún momento se quiso intentar lo mismo para proposiciones que describen los datos de la experiencia
- Ese proyecto falló, ahora lo que caracteriza la ciencia es el método (esta afirmación es muy controvertida)

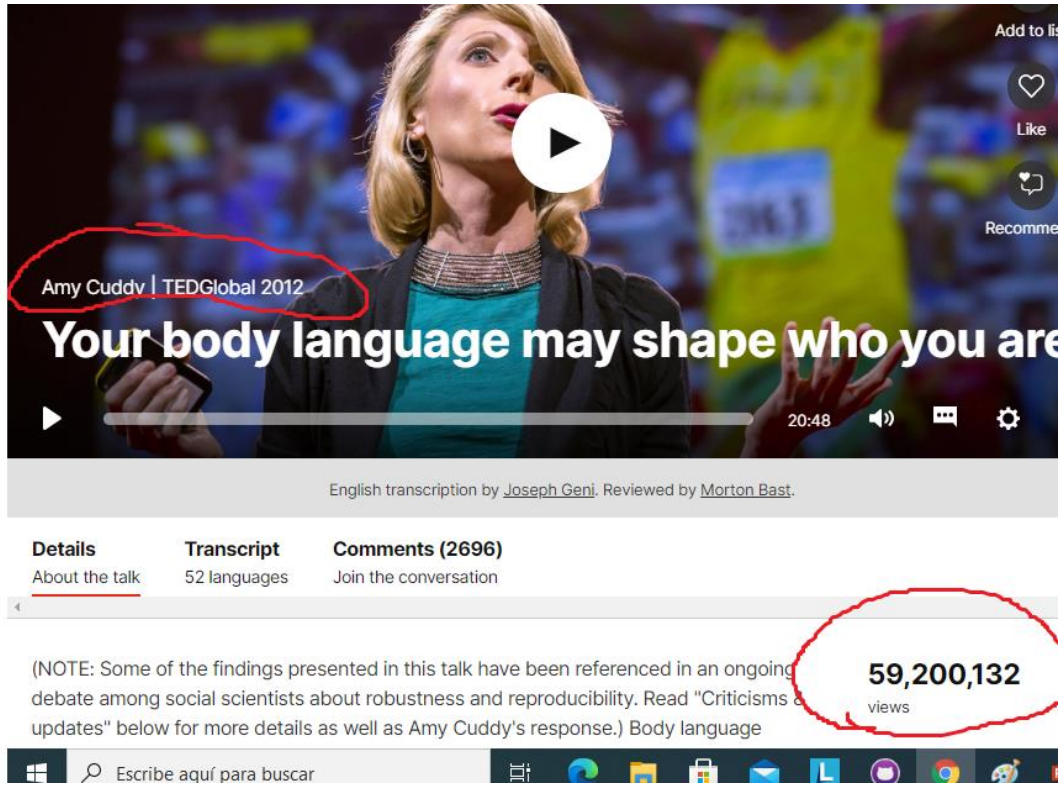
Consenso intersubjetivo

- Piensen el caso de uno de los papers que hemos visto
- Hemos aprendido que:
  - nuestro conocimiento es condicional
  - La causalidad es probabilística
  - El grado de certidumbre de nuestras afirmaciones nunca es 100% (falsos descubrimientos etc)
- ¿En qué se fonda entonces la “verdad”?
  - En un tipo de consenso intersujetivo “la posibilidad de observar lo mismo en las mismas condiciones”
  - Clave para que esto pueda funcionar es la replicabilidad

¿Por qué son empíricamente relevantes los falsos descubrimiento?

- Es un tema de sociología de la ciencia
- Ciencia es descubrimiento-> hay un premio en la “novedad”
- Sistema de incentivos:
  - Top 5-itis
  - Star market
  - Tiempos

The power pose



Research Report

aps ASSOCIATION FOR PSYCHOLOGICAL SCIENCE

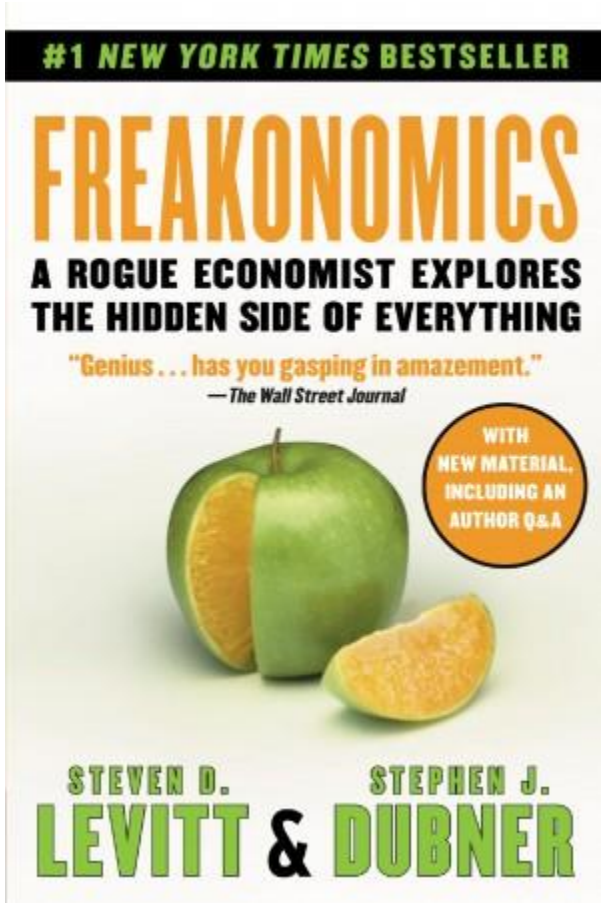
Psychological Science
21(10) 1363–1368
© The Author(s) 2010
Reprints and permission:
sagepub.com/journalsPermissions.nav
DOI: 10.1177/0956797610383437
http://pss.sagepub.com
SAGE

Power Posing: Brief Nonverbal Displays Affect Neuroendocrine Levels and Risk Tolerance

Dana R. Carney[1], Amy J.C. Cuddy[2], and Andy J. Yap[1]
[1]Columbia University and [2]Harvard University

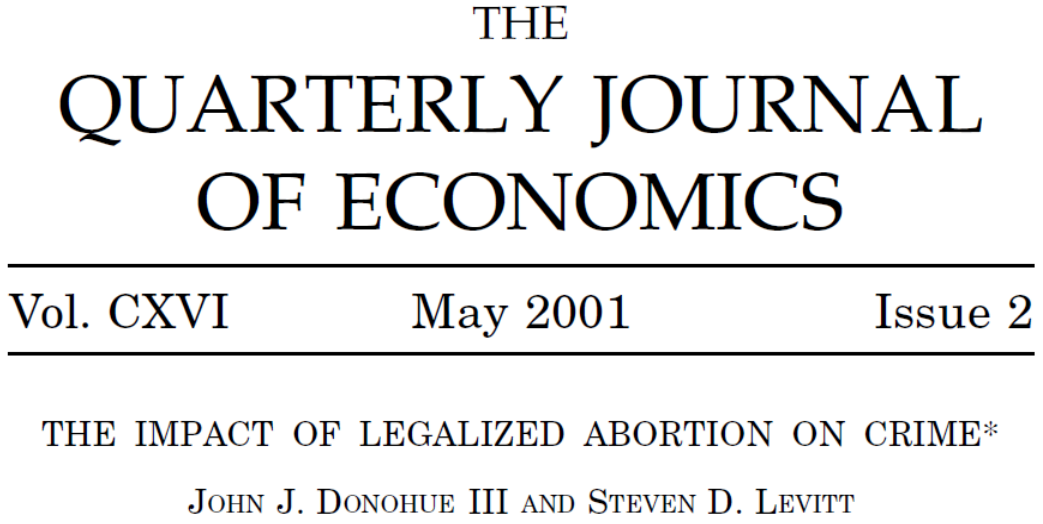
Economía

- La presencia de un mercado centralizado;
- El efecto combinado de business school, la *demanda* asiática, el boom financiero, los mecanismos de deuda para permitir el acceso;
- Los economistas estrellas



Publication bias

- Los papers que dicen “esto no funciona”, “esto no sirve” no salen publicado
- Esto puede distorsionar nuestro proceso de acumulación de conocimiento



El uso del research report

- Es posible publicare los resultados de un estudio en forma de *Research Report*
- Es muy común en las revistas de ciencia y de la salud
- Hoy otra alternativa es la pre-publicación
  - Se escribe todo el paper antes de recopilar datos
  - Se escriben los planes de análisis
  - Se formulan las hipótesis
  - La revista lo analiza antes y se compromete a publicarlo

P-hacking

When she arrived, I gave her a data set of a self-funded, failed study which had null results (it was a one month study in an all-you-can-eat Italian restaurant buffet where we had charged some people ½ as much as others). I said, "This cost us a lot of time and our own money to collect. There's got to be something here we can salvage because it's a cool (rich & unique) data set." I had three ideas for potential Plan B, C, & D directions (since Plan A had failed). I told her what the analyses should be and what the tables should look like. I then asked her if she wanted to do them.

Every day she came back with puzzling new results, and every day we would scratch our heads, ask "Why," and come up with another way to reanalyze the data with yet another set of plausible hypotheses. Eventually

BAD SCIENCE | FEB. 8, 2017

A Popular Diet-Science Lab Has Been Publishing Really Shoddy Research

By Jesse Singal

SCIENCE

Here's How Cornell Scientist Brian Wansink Turned Shoddy Data Into Viral Studies About How We Eat

Study Information

Hypotheses

H1: Trauma induces pro sociality, risk aversion, higher temporal discount and lower cognitive performance
H2: Exposure to shocks induces pro sociality, risk aversion, higher temporal discount and lower cognitive performance
H3: Expert based communication is preferred under value consensus, low epistemic uncertainty, diffused interests
H4: Deliberative is preferred under high epistemic uncertainty
H5: Negotiated is preferred under conflict of interests and value conflict
H6: There is significant fear of economic consequences of the policy response, and there is significant frustration with expert based communication strategy
H7: There will be log run behavioral change in the following direction: (positive) more cultural consumption, higher saving rate, (negative) lower earnings, less relational social capital, more risky health behavior in non CoViD19 related domains

Design Plan

Study type

Experiment - A researcher randomly assigns treatments to study subjects, this includes field or lab experiments. This is also known as an intervention experiment and includes randomized controlled trials.

Blinding

Francesco Bogliacino, francisco lupiáñez-villanueva, cristiano codagnone, Rafael Alberto Charris, Camilo Ernesto Gómez Cangrejo, F. Folkvord (Frans), Felipe Montealegre, Giuseppe A. Veltri, Giovanni Liva, and gerda reith

Description

We will conduct a longitudinal study (with four waves) in three countries t measure short run and long run behavioral change induced by the exposure to CoViD19 and response t government strategy and communication. The study includes a baseline, two interventions, and a follow up. In week two and three, we will conduct experiments using prim

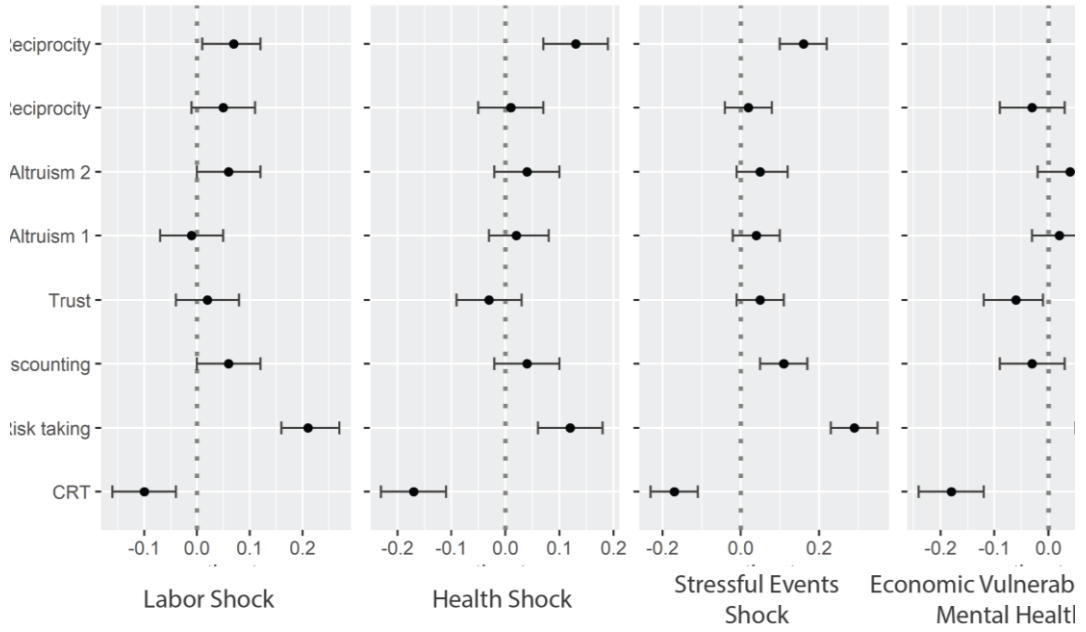
Show more

Registration type

OSF Preregistration

Date registered

April 13, 2020



Bogliacino et al. (2020)

OSF y la preregistración

Estudios experimentales y Replicabilidad

- En los estudios experimentales una gran ventaja es la replicabilidad explicita
- Además, ayuda a que se formulen hipótesis antes de recolectar los datos
- En estudios cuasi experimentales la replicabilidad de los resultados es posible en términos de *correr los mismo análisis sobre los mismos datos*, pero replicabilidad en el sentido propio del término es difícil
- El problema de la replicabilidad es que no hay incentivos
  - No es original por definición
  - No hay muchos incentivos ni de publicaciones ni de visibilidad

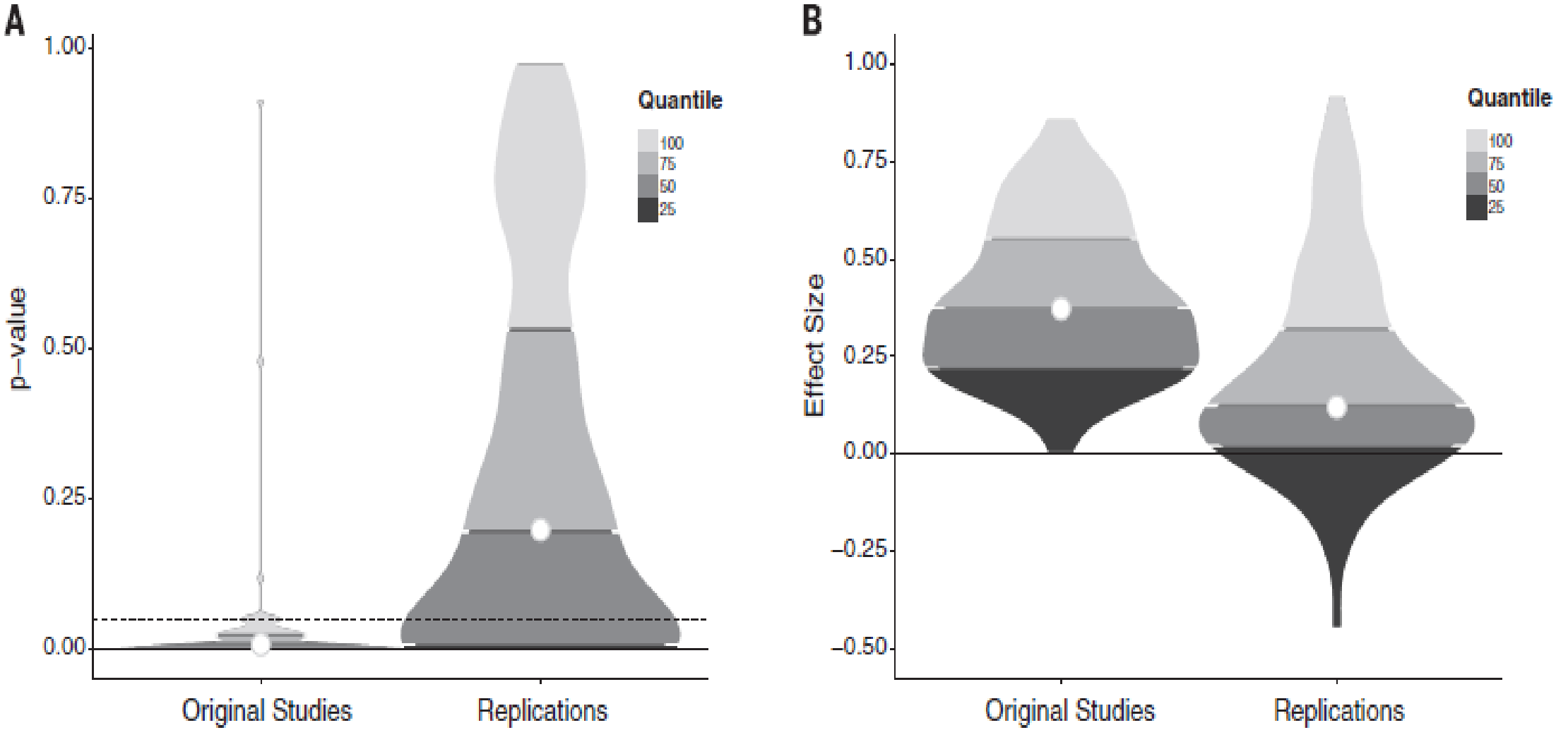
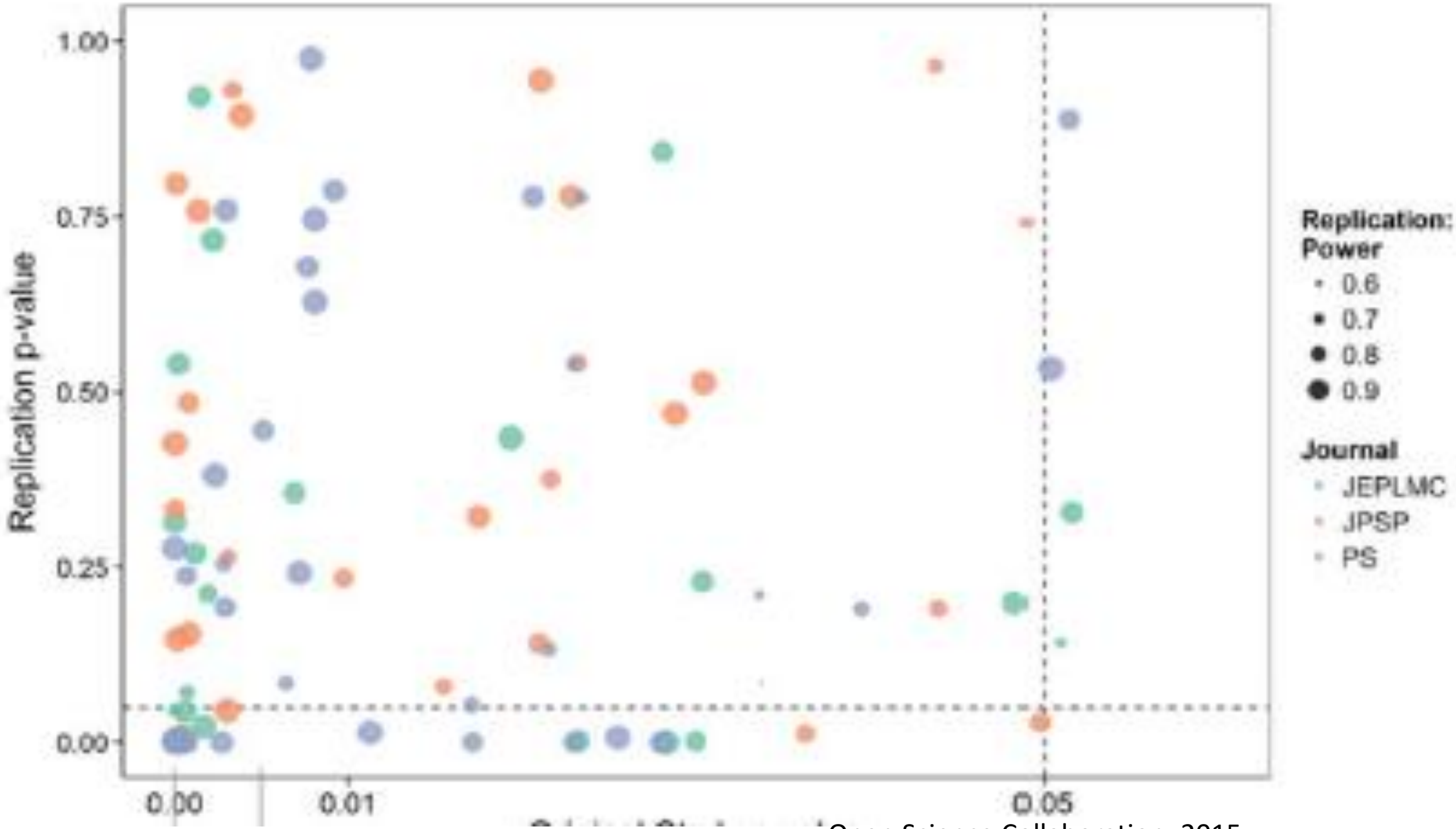


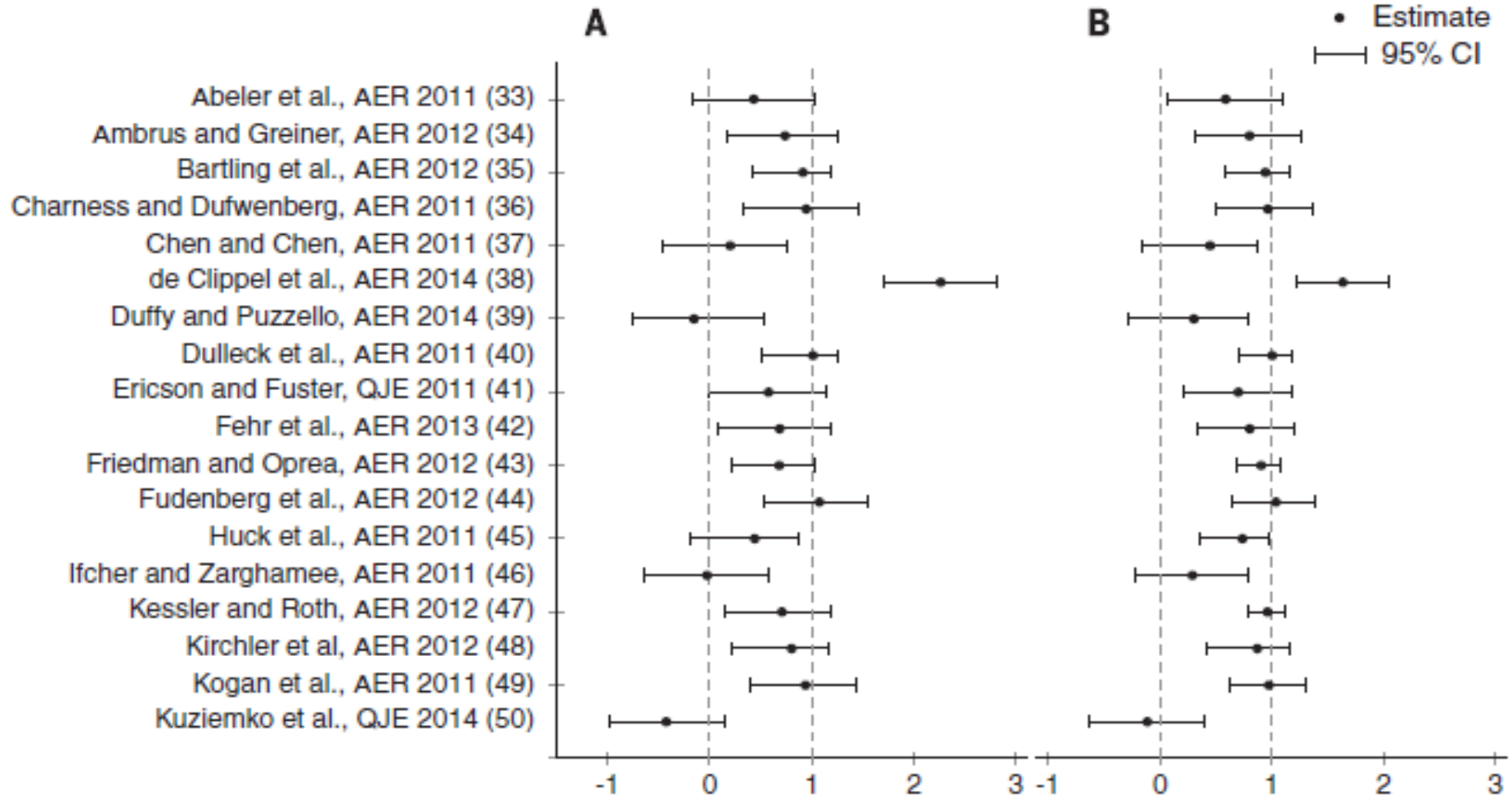
Fig. 1. Density plots of original and replication *P* values and effect sizes. (**A**) *P* values. (**B**) Effect sizes (correlation coefficients). Lowest quantiles for *P* values are not visible because they are clustered near zero.

Open Science Collaboration, 2015



Open Science Collaboration, 2015

Fig. 1. Replication results.
(**A**) Plotted are 95% CIs of replication effect sizes (standardized to correlation coefficients). The standardized effect sizes are normalized so that 1 equals the original effect size (fig. S1 shows a nonnormalized version). Eleven replications have a significant effect in the same direction as in the original study [61.1%; 95% CI = (36.2%, 86.1%)]. The 95% CI of the replication effect size includes the original effect size for 12 replications [66.7%; 95% CI = (42.5%, 90.8%)]; if we also include the study in which the entire 95% CI exceeds the original effect size, this increases to 13 replications [72.2%; 95% CI = (49.3%, 95.1%)]. AER denotes the *American Economic Review* and QJE denotes the *Quarterly Journal of Economics*. (**B**) Meta-analytic estimates of effect sizes, combining the original and replication studies. Plotted are 95% CIs ofcombined effect sizes (standardized to correlation coefficients). The standardized effect sizes are normalized as in (A) (fig. S1 shows a nonnormalized version). Fourteen studies have a significant effect in the same direction as the original study in the meta-analysis [77.8%; 95% CI = (56.5%, 99.1%)].



Source: Camerer et al. 2016

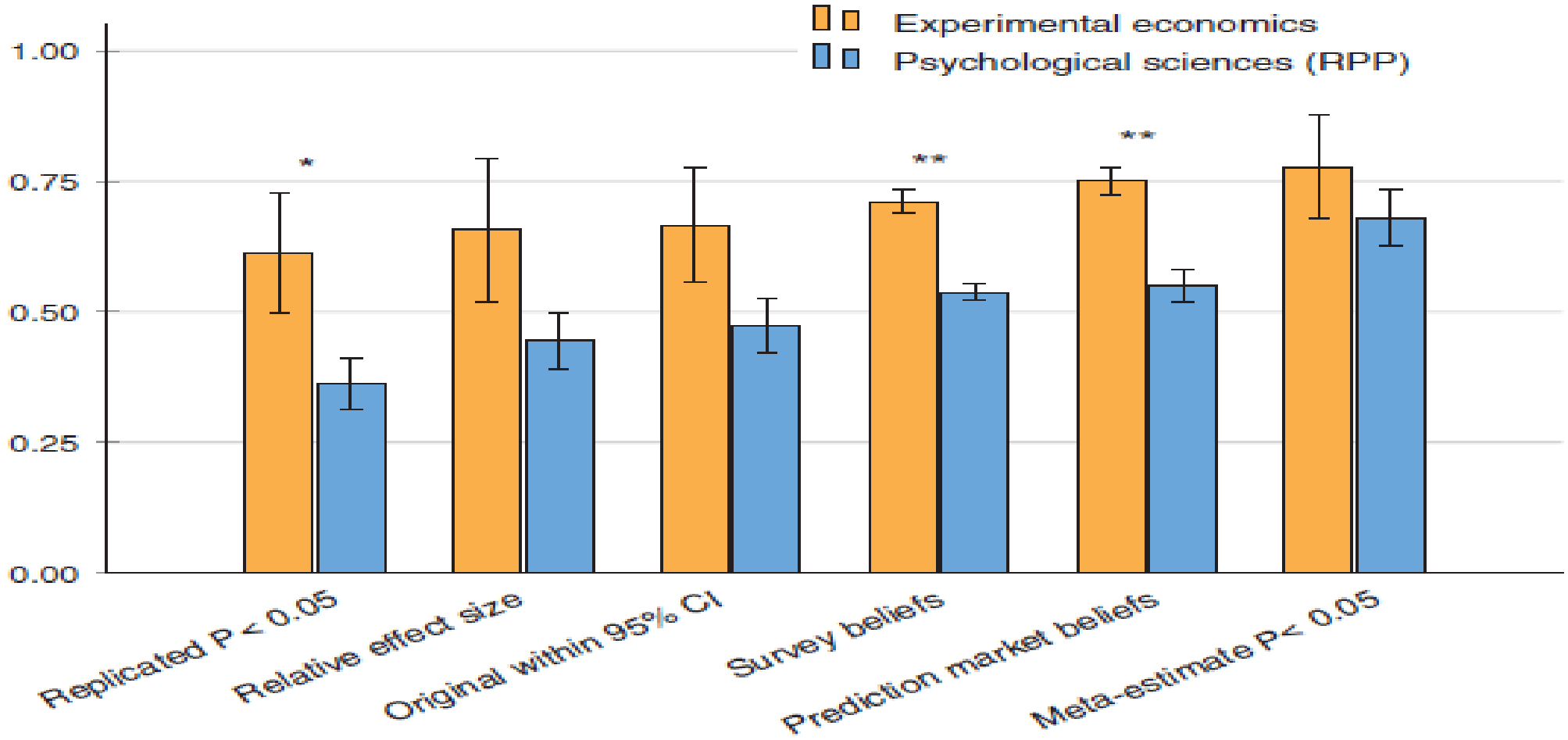


Fig. 4. A comparison of replicability indicators in experimental economics (this study) and psychological sciences (RPP). The graph shows means ± SE for replicability indicators. All six replicability indicators are higher for experimental economics; this difference is significant for three of the replicability indicators. The average difference in replicability across the six indicators is 19 percentage points. Details about the statistical tests are included in the supplementary materials. *$P < 0.05$; **$P < 0.01$.

Source: Camerer et al. 2016

Hipótesis múltiples

- Muchas veces los estudios tienen una de las siguientes tres estructuras
  - Un tratamiento (D=1 o 0) y muchos outcomes
    - Ejemplo, CCT y voy a ver si funcionan sobre aprendizaje, oferta de trabajo, ahorro, etc.
  - Un tratamiento (D=1 o 0) y un outcome, pero muchos subgrupos para ver heterogeneidad
    - Ejemplo, voy a ver si es diferente por género, edad, étnia, etc
  - Varios tratamientos y un outcome
    - Ejemplo, funciona para los niños en las escuelas repartir computadores, y mejorar la alimentación escolar, y ambos, y mejorar los docentes, y las combinaciones con los otros dos, y la jornada única, etc.
- En práctica, estoy haciendo muchas regresiones. ¿Qué impacto tiene?

Miremos un ejemplo

- Generé datos con Stata con el supuesto que el impacto del tratamiento sea cero, en diez outcomes independientes
- Tomadas separadamente, se ve que las pruebas estadísticas tienen la *size* correcta (5%)
- Qué pasa si corremos 10 regresiones?

Outcome	Falsos positivos
Y1	.048
Y2	.060
Y3	.061
Y4	.060
Y5	.065
Y6	.049
Y7	.053
Y8	.038
Y9	.048
Y10	.048

¿Cuál es la p de tener una significativa?

- El único evento que no nos sirve es que todas sean no, significativas. Ya que son independientes, esto es 0.952 X … X 0.952=0.58, entonces la probabilidad de qué por lo menos una salga significativa es del 42%
- Más precisamente, dada m hipótesis independientes y α el nivel de significancia, la prob de falso positivo es $1-(1-\alpha)^m$
- Con $\alpha = .05$

HM	Falso positivo
5	22.6%
6	26.5%
7	30.1%
8	33.6%

Dos soluciones

- índices
  - Tiene sentido para hipótesis no independientes
  - Requiere etapas:
    - Reorientar
    - Normalizar
    - Determinar ponderaciones

- Family Wise Error Rate
  - Se ajustan los p valores

Un framework para pensar el problema

H0 es	F	F	F	T	T	T	T	T	T
P value	.004	.005	.0065	.007	.04	.40	.50	.60	.70
Standard test (5%)	X	X	X	X	X	√	√	√	√

	Bonferroni	Holm	Romano Wolff
Requiere independencia	No	No	La tiene en consideración
Conservativo	Si	Depende	No
Metodología	Una etapa	Step Down	Resampling y step down

Bonferroni

H0 es	F	F	F	T	T	T	T	T	T
P value	.004	.005	.0065	.007	.04	.40	.50	.60	.70
Standard test (5%)	X	X	X	X	X	√	√	√	√
Bonferroni	X	√	√	√	√	√	√	√	√

Holm

- Controla que para la familia de M hipótesis haya una tasa de falso positivo de α (por ejemplo 5%) sin matar el poder estadístico
- Ordena las M hipótesis en término de p valor creciente (p1<p2<…<pM)

P value del test	Umbral
p1	α/M
p2	α/(M-1)
…	
pk	α/(M-k+1)
…	
pM	α

Romano Wolff

- Usa un re-muestreo para generar distribución empírica de cada una de las M hipótesis
- Calcula los estadísticos en cada submuestra
- Como Holm es step-wise:
  - Más exigente con las hipótesis más “significativas” y menos con las menos significativas
  - Se usa como umbral el percentil de la distribución empírica
- Permite calcular los p values corregidos
- Existe en Stata