Tillförlitliga forskningsresultat

ANNA DREBER ALMENBERG
är professor i nationalekonomi vid Handelshögskolan i Stockholm, Wallenberg Scholar, samt ledamot i Sveriges unga akademi, Kungliga Vetenskapsakademien och Kungliga Ingenjörsvetenskapsakademien.
Anna.Dreber@hhs.se

Det är en stor ära att få Assar Lindbeckmedaljen. Jag vill tacka mina många fantastiska medförfattare för alla gemensamma projekt – särskilt min man Johan Almenberg samt Magnus Johanneson och Thomas Pfeiffer. Något som karakteriserar dessa tre personer är att de alla är "natural born skeptics". Det är bra egenskaper inom forskningen!

I den här artikeln sammanfattar jag den forskning jag gjort med medförfattare på temat replikerbarhet och tillförlitlighet av resultat. I en direkt replikation gör man om en studie, ofta experiment, med samma eller liknande metoder och material som i originalstudien men på ett nytt och ofta betydligt större urval. I en konceptuell replikation väljer man i stället att explicit variera något för att se om det ändå går att få ett resultat som liknar det ursprungliga. Ofta mäts ett framgångsrikt replikationsutfall utifrån att replikationsresultatet är i samma riktning som originalresultatet och statistiskt signifikant ($p<0{,}05$ i ett tvåsidigt test). Det är också vanligt att studera den relativa effektstorleken – hur stor är replikationens effektstorlek relativt originalresultatet. Jag kommer i artikeln mest att skriva om hur "vi" gör eller finner något och med "vi" menar jag mina många medförfattare (där Magnus Johannesson alltid är med) och jag.

1. Vilka resultat kan vi lita på?

Tyvärr finns det många skäl att tro att en stor andel av vetenskapligt publicerade resultat inom fält såsom nationalekonomi och företagsekonomi är otillförlitliga. På gott och ont lever många av oss i en värld där vi testar nollhypoteser och alternativa hypoteser. Om vi gör en randomiserad studie där deltagare antingen får behandling eller är med i en kontrollgrupp är vi i den bästa av kausala världar. Nollhypotesen som sätts upp är att behandlingen inte har en effekt på utfallet av intresse, medan den alternativa hypotesen är att behandlingen har en effekt. Det finns många tänkbara förklaringar till de ofta falska positiva resultat som publiceras – dvs resultat där vi tror att något händer och behandlingen har en effekt när nollhypotesen i själva verket är sann. Men det finns också många potentiella lösningar. Det är naturligt att alla resultat inte visar sig vara sanna positiva (eller negativa) resultat, men vi har ett systematiskt problem om andelen falska positiva resultat konsekvent är större än vad vi borde förvänta oss givet vår accepterade nivå för statistisk signifikans. Falska resultat försvinner så småningom

Texten baseras på prisföreläsningen som arrangerades av Nationalekonomiska Föreningen den 11 mars i år.

från den vetenskapliga litteraturen, men det går att påskynda denna process samt undvika att vi initialt publicerar vissa av dessa falska resultat. Jag fokuserar här på falska positiva resultat snarare än falska negativa resultat eftersom publikationsbias främst leder till publikation av positiva resultat.

Det finns flera olika anledningar till varför vi i dag inom många fält upplever en s k replikationskris, då det visat sig att flera resultat inte håller. P-värden, eller sannolikheten att vi skulle få en effekt som är minst så stor som den observerade om nollhypotesen vore sann, är ett standardmått för att utvärdera resultat inom många kvantitativa fält. Ett resultat med ett p-värde mindre än 0,05 (och tyvärr används ibland p<0,10 för statistisk signifikans inom nationalekonomi och företagsekonomi) betecknas ofta som ”statistiskt signifikant” och tolkas som evidens för den alternativa hypotesen. P-värdet ger dock endast en bit information om sannolikheten att hypotesen som testas är sann (Ioannidis 2005). Sannolikheten att den testade hypotesen är sann beror även på den statistiska styrkan för att hitta en sann positiv effekt samt a priori sannolikheten (*prior*) att hypotesen är sann. Varningsklockor ringer därför när vi ser statistiskt signifikanta resultat från studier med låg statistisk styrka som testar ”överraskande” hypoteser. Men a priori sannolikheter är ofta subjektiva och svåra att mäta, vilket gör att de sällan diskuteras explicit kring hypotestester.

Den statistiska styrkan är ofta låg inom nationalekonomi. Ioannidis m fl (2017) går igenom mer än 6 700 studier från 159 litteraturer inom nationalekonomi och utgår från de meta-analytiska effektstorlekarna för att se vilken styrka individuella studier har att finna dessa effektstorlekar. Med de mest generösa måtten visar det sig att medianstyrkan endast är 18 procent. Zhang och Ortmann (2013) visar på något bättre men ändå ganska bedrövliga siffror för diktatorspel – ett av de absolut mest studerade spelen inom experimentell nationalekonomi – med medianstyrka på 25 procent.

Mer nyligen har David Bilén, Magnus Johannesson och jag studerat könsskillnader i diktatorspelet (Bilén m fl 2020) – något många tidigare gjort i individuella studier. I standardversionen av diktatorspelet randomiseras hälften av deltagarna till att bli diktatorer och hälften till att bli mottagare och diktatorerna får en summa pengar som de kan välja hur de vill fördela mellan sig själva och en slumpvis mottagare. Homo economicus förväntas behålla alla pengarna och inte dela med sig, men ofta ser man att många väljer att ge (och då vanligtvis hälften). Vi genomförde en meta-analys på 53 diktatorspelstudier med mer än 15 000 unika observationer för att se huruvida könsskillnader i spelet beror på om mottagaren är en annan person eller om pengarna går till välgörenhet. Vi fann att kvinnor i genomsnitt är mer generösa än män, men att effektstorleken för standardversionen där mottagaren är en annan person är liten (2,3 procentenheters skillnad) och att medianstyrkan att finna den effektstorleken endast är nio procent. Urvalsstorlekar är alltså vanligtvis på tok för små för att trovärdigt kunna studera könsskillnader i diktatorspelet. Kommer de här studierna kring statistisk styrka att ha någon effekt på framtida forskning? Oklart,

men jag hoppas så klart det. Och det finns fler problem med låg statistisk styrka – se t ex Gelman och Carlin (2014) för mer.

Kombinationen av låg statistisk styrka, testandet av spekulativa hypoteser och publikationsbias (Sterling 1959; Rosenthal 1979), där statistiskt signifikanta resultat är mer sannolika att publiceras än nollresultat, är en farlig blandning med det möjliga utfallet att en majoritet av de publicerade resultaten är falska positiva (Ioannidis 2005). En annan viktig bidragande faktor till låg tillförlitlighet av vetenskapliga resultat är de ”forskares frihetsgrader” i en analys som kan leda till att forskare medvetet eller omedvetet väljer analyser som stödjer deras hypoteser i termer av $p<0{,}05$ resultat. Dessa slags beteenden kallas *p-hacking* (Simmons m fl 2011) eller *forking* (Gelman och Loken 2013) och leder till meningslösa p-värden. För mer diskussion av dessa frihetsgrader se t ex Dreber Almenberg och Johannesson (2018). Självrapporterat beteende bland forskare tyder på stark närvaro av dessa frihetsgrader i analysen (John m fl 2012).

2. Replikationer

Systematiska replikationsprojekt är ett viktigt verktyg för att utvärdera tillförlitligheten av resultat. Tillsammans med Magnus Johannesson diskuterade jag en del av nedan beskriva projekt i *Ekonomisk Debatt* 2018 (Dreber Almenberg och Johannesson 2018). Jag går här kort genom dessa igen samt presenterar några senare projekt och jag hänvisar läsarna till den tidigare artikeln för mer information om de äldre projekten och mer utförlig diskussion om t ex forskares frihetsgrader.

I det stora replikationsprojektet inom psykologi (RPP; Open Science Collaboration 2015) genomfördes replikationer på 100 studier publicerade inom tre topptidskrifter inom kognitiv- och socialpsykologi från året 2008 (*Psychological Science*, *Journal of Personality and Social Psychology* och *Journal of Experimental Psychology: Learning, Memory, and Cognition*). Totalt var 270 forskare inblandade i projektet som leddes av psykologen Brian Nosek. Replikationerna var avsedda att ha hög statistisk styrka och hade i genomsnitt 92 procent styrka för att finna 100 procent av originaleffekten på fem procent signifikansnivå. Det här låter bra om man tror att originaleffekterna inte är kraftigt överdrivna i den mån de är sanna – om även sanna positiva resultat har överdrivna effektstorlekar kan replikationerna ha för låg styrka. Mer om det senare, men med det binära replikationsmåttet replikerar 35 av de 97 originalstudier som hade statistiskt signifikanta resultat.

I ett liknande men mindre projekt inom experimentell nationalekonomi (EERP; Camerer m fl 2016) genomförde vi (18 forskare) replikationer på 18 studier publicerade i två topptidskrifter inom nationalekonomi (*American Economic Review* och *Quarterly Journal of Economics*) under tidsperioden 2011–14. I detta projekt (till skillnad från i RPP) inkluderade vi endast studier som testade huvudeffekter (dvs ej interaktionseffekter) där deltagarna hade randomiserats till behandling eller kontroll. Replikationerna hade i

genomsnitt 92 procent styrka för att finna 100 procent av originaleffekten på fem procent signifikansnivå. Med det binära replikationsmåttet replikerar 11 av 18 studier.

Vi (24 forskare) genomförde därefter ett annat replikationsprojekt på experimentella studier inom främst nationalekonomi och psykologi (SSRP; Camerer m fl 2018). Vi granskade 21 studier publicerade i två allmänvetenskapliga topptidskrifter (*Nature* och *Science*) under tidsperioden 2010–15. Den statistiska styrkan var betydligt högre i denna studie jämfört med de två tidigare projekten just eftersom även sanna positiva effekter kan vara överdrivna i originalstudierna. Vi hade en tvåstegs design så att i steg 1 hade replikationen 90 procent styrka att finna 75 procent av originaleffekten på fem procent signifikansnivå. Om originalresultatet inte replikerade i steg 1, samlades mer data in i steg 2 så att vi med den poolade replikationsdatan hade 90 procent styrka att finna 50 procent av originaleffekten på fem procent signifikansnivå. Vi finner att 13 av 21 originalstudier replikerar i steg 2. Den relativa effektstorleken för replikationerna är 46 procent. För de 13 studier som replikerade är den 74 procent och för de åtta som inte replikerade är den noll procent.

3. Kan vi förutse vilka resultat som håller?

Ett annat verktyg för att utvärdera tillförlitligheten av resultat är att använda prognosmarknader eller enkäter för att aggregera forskares uppfattningar om studiers replikerbarhet. Prognosmarknader föreslogs för användning i forskning av Robin Hanson (1995) och användes först kring hypotesprövning av Almenberg m fl (2009).

Inspirerade av Hanson (1995) och Almenberg m fl (2009) har vi tillämpat prognosmarknader på bl a replikationsprojekten beskrivna ovan. På dessa marknader bjöd vi in forskare från olika nätverk och gav dem en summa pengar (50–100 dollar) för att köpa och sälja kontrakt med binära utfall kring huruvida resultatet replikerar eller inte. Deltagarantalet varierade från ca 30 till 200. Marknaderna var öppna i 10–14 dagar och kontrakten var generellt värda en dollar om den aktuella studien replikerade och noll dollar om den inte replikerade. Med vissa brasklappar (diskuterade av Manski 2006) tolkas priserna på dessa kontrakt som sannolikheterna marknaden tillskriver att resultaten replikerar.

Prognosmarknader användes för att förutse 41 av replikationsutfallen i RPP (Dreber m fl 2015), de 18 replikationsutfallen i EERP (Camerer m fl 2016), de 21 replikationsutfallen i SSRP (Camerer m fl 2018) samt även 44 replikationsutfall i *Many Labs* 2 och *Many Labs* 5 (Forsell m fl 2019; Ebersole m fl 2020). (*Many Labs* projekten är stora replikationsprojekt inom psykologi där klassiska och nya studier replikeras i ett flertal labb samtidigt. Liknande projekt borde genomföras i nationalekonomi!). Innan forskarna deltog på prognosmarknaderna frågade vi dem i en enkät vilken sannolikhet de tillskrev att varje hypotes skulle replikera. I den enklaste analysen säger vi

att om prognosmarknadspriset är över 50 av 100 tror marknaden att resultatet replikerar och samma sak för enkätsvaren. När vi poolar prognosdatan från dessa fem studier (N=123 för vilka vi har både prognosmarknadspriser och enkätprognoser) finner vi att marknaderna med det binära måttet på prognoser har en 72 procent korrekt prognosfrekvens (88 av 123 studier). Enkätens motsvarande siffra är 64 procent (79 av 123 studier).

Jag har också varit inblandad i andra studier där vi försöker förutse resultaten på replikationer samt nya hypoteser. I dessa projekt, där vi samarbetar med socialpsykologen Eric Uhlmann, ber vi forskare att i enkätformat förutse resultat efter att de i detalj gått igenom studiernas design och material, inklusive urvalsstorlek och mått (t ex Landy m fl 2020; Tierney m fl 2020). Vi ber forskarna att ange sannolikheten för att studien ger ett statistiskt signifikant ($p<0,05$) resultat samt standardiserad effektstorlek med riktning. Vi ställer även frågor kring konfidens, expertis och i vilken utsträckning materialet faktiskt testar hypotesen. Även i dessa studier tyder resultaten på att forskare är relativt bra på att förutse forskningsresultat.

4. Många analytiker

När en hypotes testas i ett specifikt dataset gör forskaren många olika val: statistiskt test, kontrollvariabler, utfallsmått, exkluderande av observationer, subgruppstester m m. Variationen i resultat över alla olika sätt att testa hypotesen i ett specifikt dataset ger en uppskattning för hur mycket utrymme det finns för *p-hacking*, där p-hackaren kan välja det resultat som passar hen bäst i termer av $p<0,05$. Med en förregistrerad analysplan där *p-hacking* inte är möjligt måste samma analysval göras. En metod för att förstå hur variation i resultat är beroende på olika analysval är att låta olika forskare självständigt testa samma hypotes på samma data (som i t ex Silberzahn m fl 2018). Forskares frihetsgrader och den ytterligare variansen och utrymmet för *p-hacking* beror sannolikt på vilken slags data som används för att testa hypotesen.

I projektet *Neuroimaging and Analysis Replication and Prediction Study* (NARPS; Botvinik-Nezer m fl 2020) fokuserar vi på variation i hjärnavbildningsresultat inom neurovetenskap. Våra medförfattare vid *Tel Aviv University* samlade in fMRI data från 108 deltagare som fattade beslut under osäkerhet. 70 olika analysgrupper fick detta dataset och ombads testa nio riktade hypoteser kring hjärnaktivitet. Vår huvudvariabel är andelen forskningsgrupper som rapporterar ett statistiskt signifikant resultat (ja/nej), baserat på deras egna kriterier, för varje hypotes. Vi finner stor variation i rapporterade resultat – andelen grupper som rapporterar ett statistiskt signifikant resultat varierar från sex procent till 84 procent över de nio hypoteserna. I genomsnitt rapporterar 20 procent av grupperna ett resultat som skiljer sig från majoriteten av grupperna, vilket är hög variation och tydligt visar att analytiska val påverkar rapporterade resultat. Vi satte även upp prognosmarknader där vi ser att forskare överskattar sannolikheten att

resultaten är statistiskt signifikanta. Vi genomför nu med kollegor ett liknande projekt inom finansiell ekonomi (https://fincap.academy/).

5. Andra tankar

Om vi vill öka andelen tillförlitliga resultat finns det en del lågt hängande frukter. Vi kan använda förregistrerade analysplaner i högre utsträckning, för att göra p-värden mer meningsfulla och tydliggöra vad som är bekräftande eller vad som är explorativa analyser. Explorativa analyser kan vara väldigt intressanta, men det blir fel om de framställs som bekräftande. De kan dock vara hypotesgenerande och tillåter oss att på så sätt ”upptäcka” nya resultat och fenomen som sedan kan testas i senare och förregistrerade studier. Ett problem är dock att det inte finns någon norm kring rapporterande av analyser från förregistrerade analysplaner. Ofosu och Posner (2021) jämför vad forskare ”förregistrerar” i bl a *the AEA Registry* (som många nationalekonomer använder) med vad forskarna faktiskt redovisar i sina forskningsartiklar. Ofosu och Posner finner stor variation i hur forskare följer analysplanen och redovisar avvikelser. Det här kan förbättras. Lösningen ligger nog inte på granskarna som då i detalj skulle behöva gå igenom de förregistrerade analysplanerna – bördan kan i stället ligga på författarna, som kan skriva något som liknar det följande i artikeln: ”Alla analyser nedan beskrivs i den förregistrerade analysplanen om inget annat nämns i texten. Vi delar in testerna i den förregistrerade analysplanen i primära hypotestester, robusthetstester, sekundära tester och explorativa tester; det är även så resultaten här presenteras.”

Det finns också många områden där förregistrerade analysplaner borde vara normen – för i princip alla labbexperiment, många fältexperiment och en del andra projekt. Det finns också områden då det är svårare att ha trovärdiga förregistrerade analysplaner, t ex i de fall där forskaren redan vet hur data ser ut. Då kanske något som *multiverse* analyser (se t ex Steegen m fl 2016) borde vara normen. Ett annat alternativ för icke-experimentell data från t ex SCB är att SCB endast ger en delmängd av datan som forskaren sen får ”lära känna” innan denne binder sig vid masten med en specifik analys till den resterande datan. SCB kan också möjliggöra replikationer i termer av reproducerbarhetstester eller robusthetstester i större utsträckning. Kanske kan SCB ha ett *special track* för sådana projekt? I den här artikeln har jag fokuserat på replikationer av experiment, men det finns ingen anledning att tro att problemet är större för denna sorts studier jämfört med annan data – experiment granskas just därför att de går att upprepa. Icke-experimentella studier har ofta väldigt många frihetsgrader i analysen samt fler problem med kausalitet, så att lägga mer krut på att fundera på tillförlitligheten av icke-experimenella resultat skulle vara önskvärt.

Nationalekonomi och företagsekonomi borde titta mer på psykologi som nu gör snabba framsteg. Inte bara användningen av förregistrerade analysplaner ökar där, utan även publikationer som s k *Registered Reports*,

där det som främst granskas av andra forskare är studiedesignen och de statistiska testerna *innan* forskaren har data och resultat. Det här är ett sätt att komma åt problemen med publikationsbias, då artikeln senare publiceras oavsett vad resultaten är.

Forskningsfinansiärer kan göra mer. Om forskningsfinansiärer inte vill kräva förregistrerade analysplaner kan de i alla fall uppmuntras. Jag som granskare kan sedan välja hur mycket vikt jag lägger vid huruvida forskarna har en förregistrerad analysplan eller inte. Vikten av öppen data diskuteras flitigt, vilket så klart är mycket bra, men öppen data räcker inte om koden är otillgänglig eller inte matchar den rapporterade analysen, vilket ofta verkar vara fallet. Gertler m fl går igenom 203 artiklar i nationalekonomiska topptidskrifter och finner att endast 16 procent av dessa har både rå data och användbar fungerande kod. Öppen data är heller ingen garanti för att den delade datan inte är p-hackad – den data som laddas upp på tidskriftens hemsida är kanske bara den med de variabler där $p<0{,}05$ och de andra variablerna rapporteras inte någonstans.

Det finns också anledning att tänka över vad vi kallar statistiskt signifikant. Givet bl a den låga replikationsgraden för $p<0{,}05$ resultat föreslår Benjamin m fl (2018) att flytta gränsen för statistisk signifikans till $p<0{,}005$ och fortsättningsvis hänvisa till $p<0{,}05$ resultat som *suggestive evidence*. Även om $p<0{,}005$ också är en godtycklig binär gräns, är den mer motiverad i meningen att för många olika slags a priori sannolikheter och statistisk styrka ger denna gräns en mycket lägre sannolikhet för falska positiva resultat än gränsen $p<0{,}05$.

Jag vill självklart ha mer replikationer och tror att ”hotet” om replikation och liknande övningar kan ha positiva effekter i att forskare i högre utsträckning väljer att göra studier med högre styrka och förregistrerade analysplaner. En brasklapp är dock att något kan replikera i en direkt replikation men inte nödvändigtvis i konceptuella replikationer – begränsningarna på generaliserbarheten är ofta oklara. Det finns mer att göra och många forskar nu inom det växande fältet *meta-science*, vilket jag tror är en mycket positiv utveckling och kommer att leda till många fler tillförlitliga och spännande resultat. Och det finns självklart många resultat vi redan kan lita på – och dessa resultat karakteriseras av just upprepbarhet.

REFERENSER

Almenberg, J, K Kittlitz och T Pfeiffer (2009), ”An Experiment on Prediction Markets in Science”, *PLoS ONE*, vol 4, s e8500.

Benjamin, D J m fl (2018), ”Redefine Statistical Significance”, *Nature Human Behaviour*, vol 1, publicerad online.

Bilén, D, A Dreber och M Johannesson (2020), ”Are Women more Generous than Men? A Meta-analysis”, manuskript, Handelshögskolan i Stockholm.

Botvinik-Nezer, R m fl (2020), ”Variability in the Analysis of a Single Neuroimaging Dataset by Many Teams”, *Nature*, vol 582, s 84–88.

Camerer, C F m fl (2016), ”Evaluating Replicability of Laboratory Experiments in Economics”, *Science*, vol 351, s 1433–1436.

Camerer, C F m fl (2018), ”Evaluating the Replicability of Social Science Experiments in Nature and Science”, *Nature Human Behaviour*, vol 2, s 637–644.

Dreber Almenberg, A och M Johannesson (2018), ”Vilka forskningsresultat kan vi lita på?”, *Ekonomisk Debatt*, årg 46, nr 2, s 17–29.

Dreber, A m fl (2015), ”Using Prediction Markets to Estimate the Reproducibility of Scientific Research”, *Proceedings of the National Academy of Sciences*, vol 112, s 15343–15347.

Ebersole, C R m fl (2020), ”Many Labs 5: Testing Pre-Data Collection Peer Review as an Intervention to Increase Replicability”, *Advances in Methods and Practices in Psychological Science*, vol 3, s 309–331.

Forsell, E m fl (2019), ”Predicting Replication Outcomes in the Many Labs 2 Study”, *Journal of Economic Psychology*, vol 75, s 102117.

Gelman, A och J Carlin (2014), ”Beyond Power Calculations: Assessing Type S (Sign) and Type M (Magnitude) Errors”, *Perspectives in Psychological Science*, vol 9, s 641–651.

Gelman, A och E Loken (2013), ”The Garden of Forking Paths: Why Multiple Comparisons Can Be a Problem, Even When There Is No ’Fishing Expedition’ or ’p-hacking’ and the Research Hypothesis Was Posited ahead of Time”, manuskript, Columbia University.

Gertler, P, S Galiani och M Romero (2018). ”How to Make Replication the Norm”, *Nature*, vol 554, s 417–419.

Hanson, R D (1995), ”Could Gambling Save Science? Encouraging an Honest Consensus”, *Social Epistemology*, vol 9, s 3–33.

Ioannidis, J P A (2005), ”Why Most Published Research Findings Are False”, *PLoS Medicine*, vol 2, s e124.

Ioannidis, J P A, T D Stanley och H Doucouliagos (2017), ”The Power of Bias in Economics Research”, *Economic Journal*, vol 127, s F236–F265.

John, L K, G Loewenstein och D Prelec (2012), ”Measuring the Prevalence of Questionable Research Practices with Incentives for Truth Telling”, *Psychological Science*, vol 23, s 524–532.

Landy, J F m fl (2020), ”Crowdsourcing Hypothesis Tests: Making Transparent How Design Choices Shape Research Results”, *Psychological Bulletin*, vol 146, s 451–479.

Manski, C F (2006), ”Interpreting the Predictions of Prediction Markets”, *Economic Letters*, vol 91, s 425–429.

Ofosu, G K och D N Posner (2021), ”Pre-Analysis Plans: An Early Stocktaking”, under utgivning i *Perspectives on Politics*.

Open Science Collaboration (2015), ”Estimating the Reproducibility of Psychological Science”, *Science*, vol 349, s 6251.

Rosenthal, R (1979), ”The ’File Drawer Problem’ and Tolerance for Null Results”, *Psychological Bulletin*, vol 86, s 638–641.

Silberzahn, R m fl (2018), ”Many Analysts, One Data Set: Making Transparent How Variations in Analytic Choices Affect Results”, *Advances in Methods and Practices in Psychological Science*, vol 1, s 337–356.

Simmons, J P, L D Nelson och U Simonsohn (2011), ”False-Positive Psychology Undisclosed Flexibility in Data Collection and Analysis Allows Presenting Anything as Significant”, *Psychological Science*, vol 22, s 1359–1366.

Steegen A, F Tuerlinckx, A Gelman och W Vanpaemel (2016), ”Increasing Transparency through Multiverse Analysis”, *Psychological Science*, vol 11, s 702–712.

Sterling, T D (1959), ”Publication Decisions and their Possible Effects on Inferences Drawn from Tests of Significance – or Vice Versa”, *Journal of the American Statistical Association*, vol 54, s 30–34.

Tierney, W m fl (2020), ”Creative Destruction in Science”, *Organizational Behavior and Human Decision Processes*, vol 161, s 291–309.

Zhang, L och A Ortmann (2013), ”Exploring the Meaning of Significance in Experimental Economics”, Australian School of Business Research Paper 2013 ECON 32, Sydney.