Vilka forskningsresultat kan vi lita på?

nr 2 2018 årgång 46

Inom många fält pågår det en ”replikationskris” då det visats att en stor andel av forskningsresultaten publicerade i topp-tidsskrifter inte går att upprepa när studierna görs om. Vi diskuterar här olika anledningar till den ofta höga andelen falska positiva resultat i den vetenskapliga litteraturen. Vi beskriver de replikationsprojekt vi har varit inblandade i och de försök vi har gjort att förutse replikationsresultat med hjälp av bl a prognosmarknader. Det finns anledning att vara orolig för tillförlitligheten av resultat även *inom nationalekonomi. Vi diskuterar också potentiella sätt att förbättra tillförlitligheten i forskningsresultat.*

ANNA DREBER ALMENBERG OCH MAGNUS JOHANNESSON

Anna Dreber Almenberg är professor i nationalekonomi vid Handelshögskolan i Stockholm.
anna.dreber@hhs.se

Magnus Johannesson är professor i nationalekonomi vid Handelshögskolan i Stockholm.
magnus.johannesson@hhs.se

Har vatten minne? De flesta forskare skulle säga nej, i bemärkelsen att vattenmolekyler inte kan komma ihåg vilka kemikalier de tidigare blandats med. Enligt en studie i topptidskriften *Nature* 1988 var svaret dock ja – vatten har minne (Dayenas m fl 1988). Föga förvånande visade replikationsförsök av experimenten ganska snabbt att vatten, som man kunde förvänta sig, dock inte har minne. Hur kunde en sådan studie, med resultat som var så osannolika, publiceras i en topptidskrift som *Nature*? Och var replikationsutfallen något vi hade kunnat förutse om vi hade haft vadslagning om huruvida resultaten var sanna?

Replikationer, där tidigare studier görs om för att se om resultaten håller, är en av den vetenskapliga metodens grundbultar. Vattenminneartikeln är nästan 30 år gammal men mycket tyder tyvärr på att det är vanligt att falska positiva resultat – dvs resultat som inkorrekt förkastar nollhypotesen – publiceras även i dag. Den goda nyheten är att vi nu vet mer om problemets omfattning och även har flera potentiella lösningar för att skapa mer tillförlitliga resultat.

I denna artikel beskriver vi de replikationsprojekt vi tillsammans med en mängd medförfattare har varit inblandade i inom psykologi och experimentell ekonomi och i vilken utsträckning man kan förutse replikerbarhet, dvs vilka resultat som håller. Replikationsprojekten tyder på att en stor andel av de publicerade resultaten är falska positiva resultat. Detta verkar inte vara helt överraskande för många då forskare visar sig vara ganska bra på att prediktera nivån på replikerbarheten och till viss del kan förutse vilka resultat som kommer att replikera. Det finns inget som tyder på att våra resultat för experimentella studier skulle vara en övre gräns på andelen falska positiva resultat i den ekonomiska litteraturen – tvärtom. En hög andel falska positiva resultat behöver inte bero på medvetet fuskande forskare. Den största

Vi är tacksamma för finansiellt stöd från Handelsbankens forskningsstiftelser, Knut och Alice Wallenbergs stiftelse (A D A är Wallenberg Academy Fellow) samt Riksbankens Jubileumsfond. Vi tackar också våra många medförfattare på dessa projekt, bl a Johan Almenberg, Adam Altmejd, Eskil Forsell, Emma Heikensten och Siri Isaksson. Vi tackar också Lina Aldén för kommentarer på manuskriptet.

boven är kanske i stället de olika frihetsgraderna forskare har i den statistiska analysen som leder till missvisande p-värden. Dessa frihetsgrader är större i studier baserade på observationsdata än i experimentella studier. Även om inte just ekonomer har hävdat att vatten har minne så finns det således många skäl till oro. Men det finns också uppenbara lösningar som kan förbättra tillförlitligheten av (ekonomiska) forskningsresultat.

1. Varför så många falska positiva resultat?

Många faktorer bidrar till att skapa en hög andel falska positiva resultat i den vetenskapliga litteraturen – en andel som inom vissa fält mycket väl kan vara en majoritet av publicerade resultat (Ioannidis 2005). Det finns t ex många kända fall av forskare som fabricerar data (Callaway 2011). Men vetenskapliga bedragare som Andrew Wakefield (som hävdade att MMR-vaccinet orsakade autism), Diedrik Stapel (som fabricerade data i en rad topp-publikationer inom psykologi) och Paolo Macchiarini är med stor sannolikhet inte huvudanledningen till de många falska positiva resultaten.

En del av den höga andelen falska positiva resultat beror på gränsen för vad som anses vara ett statistiskt signifikant resultat och på bristande statistisk styrka (*power*) i empiriska studier. Inom många vetenskaper finns det en norm att betrakta ett p-värde under 0,05 som ett statistiskt signifikant resultat, vilket då tolkas som ett starkt stöd för att den testade hypotesen är sann. Ett p-värde på 0,05 ger dock inte något starkt stöd för att hypotesen är sann om inte apriori-sannolikheten (*priorn*, den initiala sannolikheten) är mycket hög (se mer om apriori-sannolikheten nedan). Även utan medveten eller omedveten snedvridning av resultaten kan vi därför förvänta oss en ganska hög andel falska positiva resultat för studier med ett p-värde nära 0,05 (Benjamin m fl 2017). Inom nationalekonomi är det t o m vanligt att rapportera $p<0{,}10$ som statistiskt signifikant, vilket ytterligare ökar andelen falska positiva resultat.

Statistisk styrka är sannolikheten att hitta ett statistiskt signifikant resultat givet att hypotesen är sann. Statistisk styrka beror på studiens storlek (antal observationer), variansen i data och den förväntade effektstorleken om hypotesen är sann. En låg statistisk styrka ökar risken att ett statistiskt signifikant resultat är ett falskt positivt resultat och det ökar risken att effektstorleken överskattas för sanna positiva resultat (Gelman och Carlin 2014; Leamer 1983). En nyligen publicerad studie av Ioannidis m fl (2017) visar också att det finns anledning att vara orolig för bristande statistisk styrka inom empirisk nationalekonomi. Ioannidis m fl undersöker 159 olika områden i empirisk nationalekonomi med fler än 6 700 studier. De finner att en majoritet av dessa studier har för låg styrka (definierad till att ha mindre än 80 procent styrka att kunna hitta den sanna effektstorleken). Medianstyrkan är endast 18 procent, vilket är lägre än de 21 procent som tidigare rapporterats för neurovetenskap (Button m fl 2013). Ioannidis m fl finner att de rapporterade effekterna är minst dubbelt så stora som

de sanna effekterna för en majoritet av områdena och minst fyra gånger så stora som de faktiska effekterna för minst en tredjedel av områdena. Det har även visats att låg styrka är ett problem i experimentell ekonomi, där det rimligen är enklare att kontrollera statistisk styrka då urvalsstorlek är något forskaren ofta styr över. Zhang och Ortmann (2013) rapporterar att bland experiment på diktatorspelet är medianstyrkan endast 25 procent medan genomsnittlig styrka är 38 procent.

En annan variabel som är relevant när vi ska utvärdera om ett statistiskt signifikant resultat är sant eller inte är apriori-sannolikheten – den initiala sannolikheten att den testade hypotesen är sann. Att testa hypoteser med låg initial sannolikhet att vara sanna ökar risken för att hitta falska positiva resultat (Ioannidis 2005) och apriori-sannolikheten borde egentligen vägas in tillsammans med statistisk styrka och p-värde för att tolka trovärdigheten i ett publicerat resultat. Apriori-sannolikheter är dock subjektiva och svåra att få tillgång till (även om en del apriori-sannolikheter som t ex sannolikheten att vatten har minne rent objektivt borde vara väldigt låg). Prognosmarknader och andra informationsaggregationsverktyg kan vara ett sätt att uppskatta apriori-sannolikheter: kemister som slår vad med varandra hade troligen kunnat förutse de misslyckade replikationsresultaten för vattenminne-studien.

Falska positiva resultat publiceras också på grund av s k *skrivbordseffekter* eller *rapporteringsbias* (Rosenthal 1979) och *publikationsbias* (Sterling 1959). Rapporteringsbias orsakas av forskare som väljer att inte kommunicera nollresultat medan publikationsbias orsakas av tidskrifter som väljer att inte publicera nollresultat. Franco m fl (2014) finner stöd för båda fenomenen bland de 221 studier genomförda på *Time-sharing Experiments in the Social Sciences* (TESS). De finner att starka resultat relativt nollresultat oftare dokumenteras i *working papers* och oftare publiceras.

En av de troligen viktigaste källorna till falska positiva resultat i den vetenskapliga litteraturen har att göra med de olika frihetsgraderna inom forskningen som leder till att forskarna själva medvetet eller omedvetet snedvrider resultaten för att hitta statistiskt signifikanta resultat (Simmons m fl 2011; Gelman och Loken 2013). Det finns flera olika relaterade begrepp kring detta i litteraturen som vi kort beskriver nedan. Att fiska efter resultat (i litteraturen benämns detta som *fishing*) innebär att forskaren gör många olika analyser utan någon specifik hypotes och letar efter signifikanta samband och där statistiskt signifikanta resultat rapporteras medan insignifikanta ofta ”glöms bort” (och ingen justering görs av p-värdet för multipla tester).

P-hacking (Simmons m fl 2011) hänvisar till en process då forskare kan ha en specifik hypotes och sedan aktivt försöka hitta statistisk signifikans genom att t ex testa olika sorters regressioner eller andra statistiska test, inkludera eller exkludera observationer eller kontrollvariabler (där det ofta finns mycket godtycklighet kring vad som är en relevant kontrollvariabel) för att få ett p-värde under den relevanta gränsen (oftast $p<0{,}05$), analysera

många mått men bara rapportera dem med resultat som ger p<0,05, samla in och analysera många grupper men bara rapportera dem med resultat som ger p<0,05, etc. *P-hacking* innebär således ett avsiktligt sökande efter statistiskt signifikanta resultat där p-värden över den relevanta gränsen för statistisk signifikans hackas ner till att hamna under.

The garden of forking paths (Gelman och Loken 2013) är ett fenomen där en forskare kan ha avsett att testa en mycket specifik hypotes utan att ha specificerat den exakta analysen. Genom att låta analysen bero på resultaten vandrar forskaren ner för en av många möjliga vägar och hamnar med ett statistiskt signifikant resultat där p-värdet är meningslöst. Om vi t ex vill studera hur en förändring i inkomstskatten påverkar arbetsutbud med hypotesen att en ökning leder till lägre utbud finns det väldigt många förgreningar. En effekt skulle kunna finnas hos män men inte kvinnor, med förklaringen att män reagerar mer på incitament. Eller hos kvinnor men inte män, med förklaringen att kvinnor i högre utsträckning än män kanske föredrar familjetid när priset förändras. Eller hos både kvinnor och män. Eller hos yngre – de har inte redan vant sig vid en viss arbetstid/insats. Eller hos äldre – de har gjort fler val och vet mer om vad de gillar. Vilket slags test ska göras, parametriskt eller icke-parametriskt? Vilka kontrollvariabler ska inkluderas? Hur definieras förändring i arbetsutbudet, är det binärt eller kontinuerligt? Och så vidare. Det räcker med att forskarna gör ett enda test (och blir nöjda med svaret så att de inte behöver vandra vidare) – det är ändå *forking* så länge testet inte är förspecificerat. *Fishing*, *p-hacking* och *forking* är liknande fenomen, men begreppen *fishing* och *p-hacking* innebär en mer medveten process av forskaren för att snedvrida resultaten medan *forking* kan vara en mer omedveten process där forskaren själv inte är medveten om att alla val som görs under analysprocessen systematiskt snedvrider resultaten. Vid *forking* är det möjligt att forskaren bedrar sig själv kring resultatet minst lika mycket som andra blir bedragna.

2. Replikerbarhet

Det finns åtskilliga sorters replikationer.[1] En slags replikation handlar om att bekräfta att en artikels rapporterade resultat kan återskapas med exakt samma metoder och data. Detta är egentligen ett test av rena felaktigheter i rapporterade resultat i publicerade studier. Åtskilliga studier har visat att statistiska resultat inom nationalekonomi inte alltid kan återskapas (Dewald m fl 1986; McCullough och Vinod 2003; McCullough m fl 2006; Chang och Li 2015). Ett problem orsakas av otillgänglig data och kod, men även när dessa är tillgängliga går alla resultat inte att återskapa.

En annan kategori av replikationer innebär att ny data samlas in för att testa ett tidigare resultat på ett nytt urval. Denna kategori kan i sin tur delas in i direkta och konceptuella replikationer. En direkt replikation innebär att replikationen genomförs på (ibland nästintill) exakt samma sätt som origi-

[1] Se Open Science Collaboration (2015) för mer diskussion.

nalstudien, helst med det material som originalstudien använde. En konceptuell replikation innebär i stället att replikationen testar samma hypotes som originalstudien men med något andra metoder för att se om resultatet håller även under något annorlunda (kontrollerade) omständigheter. Vi fokuserar i denna artikel främst på direkta replikationer.

En komplikation i direkta replikationer är att replikationspopulationen sällan är identisk till originalpopulationen. I stället försöker man ofta ha en så lik population som möjligt (om originalstudien t ex genomfördes på amerikanska studenter försöker man ofta genomföra även replikationen på amerikanska studenter). Om behandlingseffekten (*treatment effect*) – det som studier avser testa genom att t ex randomisera olika grupper till olika behandlingar – varierar mellan populationer kommer detta att leda till ökad variation i resultat mellan originalstudierna och replikationsstudierna men det skulle inte leda till någon systematisk bias i replikationernas uppskattade effektstorlek. De flesta replikationsstudier vi diskuterar här fokuserar på att replikera just behandlingseffekter.[2]

Det finns inte ett unikt universellt accepterat kriterium för att definiera en lyckad replikation, dvs om ett resultat håller eller inte när studien replikeras (Gelman och Stern 2006; Cumming 2008; Verhagen och Wagenmakers 2014; Open Science Collaboration 2015). Den binära definitionen ”signifikant effekt i samma riktning som originalstudien” är den mest flitigt använda och var den primära replikationsindikatorn i det stora replikationsprojektet inom psykologi (*Reproducibility Project Psychology*, RPP) och replikationsprojektet inom nationalekonomi (*Experimental Economics Replication Project*, EERP). Med denna indikator söker replikationen stöd för originalstudiens hypotes med samma statistiska test som användes i originalstudien och studien anses vara replikerad om resultatet går i samma riktning som i originalstudien och är statistiskt signifikant på $p<0{,}05$ med ett dubbelsidigt test.

Relativ effektstorlek har också använts som komplement till det binära måttet. Korrelationskoefficienten (r) används här ofta som en standardiserad effektstorlek för att jämföra effektstorlekar mellan originalstudier och replikationer. Korrelationskoefficienten kodas som positiv för originalstudien oavsett effektens tecken och replikationseffektstorleken kodas som positiv om den är i samma riktning som originalstudien och negativ om den är i motsatt riktning. Den relativa effektstorleken ger ett kontinuerligt mått på graden av replikation. Om det inte finns någon systematisk bias i publicerade resultat ska den relativa effektstorleken i replikationer i genomsnitt vara 100 procent (dvs det bör vara lika vanligt att replikationerna hittar en större effekt som att de hittar en lägre effekt). Om den relativa effektstorleken är under 100 procent kan det både bero på falska positiva resultat och på att effektstorlekarna i sanna positiva resultat är överskattade. Det finns

[2] *Many Labs*-projekteten som replikerar specifika psykologistudier i en mängd labb samtidigt har inte funnit stöd för att det skulle vara mycket systematisk variation i genomsnittliga behandlingseffekter över olika urval (Klein m fl 2014; Ebersole m fl 2016).

även andra replikationsmått som har föreslagits i litteraturen: se Camerer m fl (2018) för mer diskussion.

3. Replikationsprojekt

Reproducibility Project Psychology (RPP)

RPP replikerade 100 studier publicerade i tre topptidskrifter inom kognitiv- och socialpsykologi under år 2008: *Psychological Science*, *Journal of Personality and Social Psychology* och *Journal of Experimental Psychology: Learning, Memory, and Cognition*. 488 artiklar publicerades i dessa tidskrifter under 2008. 158 av dessa valdes på basis av ett antal kriterier ut för replikation under projektperioden. Av dessa 158 artiklar valdes 111 av en replikationsgrupp (vilket ledde till 113 replikationer eftersom två artiklar hade två replikationer). 100 av dessa 113 replikationer var genomförda vid projektets slutdatum. En majoritet av artiklarna involverade experiment med behandlingseffekter men några var korrelationella. Flera artiklar rapporterade mer än en studie; i de fallen valdes oftast den sista studien ut.

Replikationsförsöken i RPP involverade förregistrerade replikationsrapporter där replikationen i detalj jämfördes med originalstudien med avseende på metoder och material samt analys. Dessa rapporter skickades efter intern granskning till originalförfattarna för återkoppling. Replikationsrapporterna hade i genomsnitt 92 procent statistisk styrka att finna 100 procent av originalstudiens rapporterade effektstorlek på en signifikansnivå på fem procent.

Endast en tredjedel av resultaten gick att replikera enligt den primära replikationsindikatorn. 97 av originalresultaten rapporterade statistiskt signifikanta resultat ($p<0{,}05$).[3] Endast 35 (36 procent) av replikationerna fann en signifikant effekt i samma riktning som originalresultatet. Det finns också indikationer på att resultat inom kognitiv psykologi kunde replikeras i större utsträckning än resultat inom social psykologi och att huvudeffekter till skillnad från interaktionseffekter var mer tillförlitliga.

Experimental Economics Replication Project (EERP)

Tillsammans med medförfattare (Camerer m fl 2016) genomförde vi även ett liknande systematiskt men storleksmässigt mindre replikationsprojekt inom nationalekonomi. Vi identifierade 18 replikationer av labbexperiment publicerade i *American Economic Review* och *Quarterly Journal of Economics* under perioden 2011–14. Till skillnad från i RPP inkluderade vi här endast studier som testade huvudeffekter där olika deltagare randomiserades till olika grupper. Replikationerna hade i genomsnitt 92 procent statistisk styrka att finna 100 procent av originalstudiens rapporterade effektstorlek på signifikansnivå på fem procent.

Det gick att replikera 11 av 18 studier, dvs 61 procent av resultaten, enligt

[3] Fyra av dessa resultat hade $p<0{,}06$.

definitionen med signifikant resultat i samma riktning som originalstudien. Detta var signifikant lägre än den genomsnittliga statistiska styrkan på 92 procent ($p<0{,}001$). Den genomsnittliga relativa effektstorleken var 66 procent om vi tar genomsnittet av den relativa effektstorleken i varje studie och 59 procent om vi delar den genomsnittliga replikationseffekten med den genomsnittliga originaleffekten (som RPP gjorde och rapporterade siffran 49 procent).

En brist i både RPP och EERP är att den statistiska styrkan i replikationerna beräknades utifrån de observerade effektstorlekarna i originalstudierna. På grund av publikations- och rapporteringsbias är det troligt att effektstorlekarna är överskattade i originalstudierna även för sanna positiva resultat. Det gör att beräkningen på statistisk styrka idealiskt bör ta hänsyn till detta och ha hög statistisk styrka att hitta lägre effekter än de som observerades i originalstudierna. De observerade replikationsnivåerna, dvs andelen signifikanta resultat i samma riktning som i originalstudierna, i RPP och EERP kan därför betraktas som en undre gräns på andelen sanna positiva resultat i dessa urval.

4. Förutsägbarhet

P-värdet i originalartikeln

Både RPP och EERP studerade korrelationen mellan originalstudiernas p-värde och replikationsutfallen. P-värden i originalartiklarna borde korrelera negativt med replikerbarhet eftersom risken av falska positiva resultat ökar med p-värdet och båda projekten finner negativa korrelationer mellan originalstudiernas p-värden och replikationsutfall (Spearmankorrelation –0,33 för RPP och –0,57 för EERP). Detta indikerar att en sänkning av p-värdesgränsen för vad som anses vara ett statistiskt signifikant resultat är ett effektivt och enkelt sätt att minska andelen falska positiva resultat i publicerade studier; vilket vi återkommer till i diskussionsavsnittet nedan.

Prognosmarknader

Prognosmarknader används för att aggregera privatinformation. På dessa marknader handlar deltagare kontrakt med väldefinierade utfall där marknadens prediktion är samma som prediktionen av en enskild deltagare som innehar all information. Med vissa brasklappar (diskuterade av Manski 2006) kan priset på ett kontrakt med ett binärt utfall tolkas som sannolikheten marknaden tillskriver utfallet. Prognosmarknader har därför använts inom områden såsom sport och politik (Wolfers och Zitzewitz 2006; Arrow m fl 2008) och föreslogs för användning i forskning år 1995 av Robin Hanson (Hanson 1995) och implementerades sedan av Almenberg m fl (2009).

Vi använde prognosmarknader för att förutse 44 RPP-studiers replikationsutfall (Dreber m fl 2015). Varje deltagare fick av oss 100 dollar som

kunde användas för att investera i kontrakt som representerade replikationerna och det binära replikationsmåttet innan replikationsutfallet var känt. Kontrakten var värda en dollar om studien replikerades och annars noll. Marknaderna var öppna i två veckor vid två tillfällen (år 2012 och 2014) med 23 respektive 21 studier per tillfälle och 47 respektive 45 deltagare. Deltagarna var främst forskare inom psykologi.

Vi använde även prognosmarknader för att förutse utfallen i EERP (Camerer m fl 2016). Innan vi genomförde de 18 replikationerna hade vi en omgång marknader öppna under tio dagar där 97 deltagare, främst nationalekonomer, fick 50 dollar var att handla med. I både RPP och EERP marknaderna fick deltagarna information om både originalstudierna och replikationerna. Våra marknader använder Hansons logaritmiska *market maker* (Hanson 2007) som erbjuder deltagarna möjligheten att köpa och sälja kontrakt vid alla tillfällen och justerar priset efter varje handel. Prognosmarknadspriset kan tolkas som sannolikheten marknaden tillskriver att en originalstudie kan replikeras. Vi frågade även deltagarna på prognosmarknaderna om vilken sannolikhet de tillskrev att en hypotes skulle replikera och hur bekanta de var med området, innan de deltog på marknaden.

Dreber m fl (2015) fann för RPP ett genomsnittligt marknadspris på 55 procent (från 13 procent till 88 procent), vilket tydde på att ca hälften av de 44 resultaten förväntades vara replikerbara. 41 av 44 replikationer slutfördes, där 16 (39 procent) gick att replikera och 25 (61 procent) inte gick att replikera enligt den binära replikationsindikatorn. Även om det genomsnittliga marknadspriset var något högre än andelen studier som kunde replikeras var deltagarna ganska bra på att prediktera nivån på andel replikationer. Deltagarna kunde också till viss del förutse vilka studier som var mer eller mindre sannolika att vara replikerbara, med en korrelationen mellan marknadspris och replikationsutfall på 0,42.

Camerer m fl (2016) fann ett genomsnittligt marknadspris på 75 procent för EERP, vilket kan tolkas som att tre fjärdedelar av de 18 studierna förväntades vara replikerbara. Endast 11 av 18 av studierna (61 procent) gick att replikera enligt den binära replikationsindikatorn. Korrelationen mellan marknadspris och replikationsutfall (0,30) var på en liknande nivå som i studien på RPP.

Dreber m fl uppskattar också sannolikheten att hypotesen som testas är sann, också kallad *the positive predictive value* (PPV) (Ioannidis 2005; Button m fl 2013). En kombination av marknadspriser, statistisk styrka och statistisk signifikans i originalstudier och replikationer tillåter en uppskattning av PPV vid tre olika tillfällen i den vetenskapliga processen: innan originalstudiens resultat är känt, efter originalresultatet och efter replikationsutfallet. Detta tillåter oss att förstå mer av dynamiken kring vilka slags hypoteser det är som forskare testar samt hur mycket vi lär oss av en enskild studie och av replikationer. Resultaten visar att apriori-sannolikheter för de 44 studierna varierar från 0,7 procent till 66 procent med en median (ett genom-

snitt) på 8,8 procent (13 procent).[4] Detta låga nummer skulle kunna reflektera topp-psykologitidsskrifters benägenhet att publicera överraskande resultat, där överraskande betyder positiva resultat från relativt osannolika hypoteser. Efter det positiva resultatet i den första studien ökar PPV och varierar från tio procent till 97 procent med ett medianvärde (genomsnitt) på 56 procent (57 procent). Detta resultat tyder på att ca hälften av de statistiskt signifikanta resultaten publicerade i dessa topptidsskrifter är falska positiva resultat. Vi uppskattar även den posteriöra sannolikheten för de 41 studier som replikerades. Bland de hypoteser som replikeras (16 studier) varierar sannolikheten från 93,0 procent till 99,2 procent med en median (ett genomsnitt) på 98 procent (97 procent). För de misslyckade replikationerna (25 studier) är i stället variationen i sannolikhet mellan 0,1 procent och 80 procent med ett medianvärde (genomsnitt) på 6,3 procent (15 procent). Dessa resultat visar hur prognosmarknader på replikationer med hög statistisk styrka kan ge viktiga insikter. Ett publicerat resultat med p-värde under 0,05 misstolkas ofta som att hypotesen har 95 procent sannolikhet att vara sann. Här visar vi att vi kan nå sådana höga sannolikheter, men bara om utfallet från en replikation med hög statistisk styrka stöder originalresultat. Givet informationsvärdet av replikationer med hög statistisk styrka kan man således fråga sig om incitamenten för replikationer är tillräckliga inom ekonomisk forskning.

5. Diskussion

I denna artikel har vi främst diskuterat replikationer för experimentell forskning inom psykologi och nationalekonomi som indikerar en ganska hög andel falska positiva resultat. Speciellt bör replikationsresultaten för experimentell ekonomi tolkas försiktigt eftersom de baserades på endast 18 replikationer. En viktig fråga är hur stort problemet med falska positiva resultat kan förväntas vara bland icke-experimentella studier inom nationalekonomi. Vår hypotes är att problemet med falska positiva resultat är än större bland icke-experimentella studier (inklusive s k naturliga experiment). Anledningen till detta är att forskarnas frihetsgrader att systematiskt snedvrida resultaten är större för den här typen av studier. Brodeur m fl (2016) hittar stöd för denna hypotes när de studerar fördelningen av p-värden i publicerade artiklar i de tre topptidskrifterna *American Economic Review*, *Journal of Political Economy* och *Quarterly Journal of Economics*. De hittar tydliga tecken på *p-hacking* och att problemet är större i icke-experimentella studier.

Det finns flera sätt att öka andelen sanna resultat i den vetenskapliga litteraturen. Högre statistisk styrka och större urvalsstorlekar är viktigt för att öka trovärdigheten av originalstudier. Att genomföra replikationer är också viktigt. Vetskapen om att det är troligt att en studie kommer att

[4] Johnson m fl (2017) skattade också apriori-sannolikheter för originalhypoteserna testade i RPP och finner liknande låga apriori-sannolikheter.

replikeras kan ha en hämmande effekt på *p-hacking*. En annan viktig källa för ökad trovärdighet är förregistrering av analysplaner för att minska frihetsgraderna i analysen och därmed minska *fishing*, *p-hacking* och *forking* (se även Casey m fl 2012 för diskussion av analysplaner inom utvecklingsekonomi). Förregistrering av analysplaner borde bli normen inom empirisk nationalekonomi. Förregistrering hindrar inte att explorativ analys kan göras efter datainsamlingen, men den gör det tydligt att analysen gjordes efter att forskaren har sett datan vilket ger resultaten mindre trovärdighet (men den explorativa analysen kan generera nya hypoteser som sedan testas mer rigoröst i framtida studier). Förregistrering av analysplaner kan också användas för att motarbeta publikationsbias. Transparens och öppenhet är också av stor vikt för att öka trovärdigheten av publicerade resultat (Miguel m fl 2014; Nosek m fl 2015).

Ett annat effektivt sätt att minska andelen falska positiva resultat är att sänka gränsen för vad som anses vara ett statistiskt signifikant resultat. En stor grupp samhällsvetenskapliga forskare (varav vi är två) föreslog nyligen att p-värdesgränsen för statistisk signifikans för nya fynd borde sänkas från 0,05 till 0,005 (Benjamin m fl 2017). Förutom att minska andelen falska positiva resultat skulle detta dessutom medverka till att urvalsstorleken i vetenskapliga studier ökar eftersom ett större urval behövs för att ha bibehållen statistisk styrka vid test på 0,005 nivån. För att projekt med sådana urval ska bli av kan forskning i större grupper (*team science*) (Klein m fl 2014; Munafò m fl 2017) där flera grupper samarbetar på liknande frågeställningar vara en lösning.

För experimentella studier inom nationalekonomi tror vi att en kombination av förregistrering och en sänkt gräns för vad som anses vara ett statistiskt signifikant resultat tillsammans skulle vara effektiva åtgärder för att väsentligt minska andelen falska positiva resultat. Att det regelbundet genomförs replikationer är också en viktig byggsten för att säkerställa tillförlitligheten av publicerade resultat (i alla fall för labbexperiment; för fältexperiment är det betydligt svårare).

Vi rekommenderar dessa åtgärder även för icke-experimentella studier, men tror att utmaningen att nå en hög trovärdighet i publicerade resultat är större inom detta område. Det är svårare att genomföra direkta replikationer, det finns metodproblem kring att trovärdigt skatta kausala effekter och det är inte säkert att förregistrering är lika effektivt eftersom det ofta är svårare att verifiera att forskarna inte har haft tillgång till data innan analysplanen registrerades. Men även för icke-experimentella studier tycker vi att förregistrering av analysplaner är ett viktigt steg på vägen. Förregistrering och en ökande medvetenhet om problemen med *p-hacking* kan också leda till en förändring av normerna inom professionen, där vi lär ut en mer korrekt vetenskaplig metodik till framtida generationer av forskare.

Kommer nationalekonomisk forskning att följa dessa rekommendationer? Det beror nog delvis på hur viktigt forskare faktiskt tycker att det är att hitta sanna resultat, snarare än de resultat som är lättast att publicera.

Mycket tyder på att sanningssökande inte alltid är det som maximeras i den vetenskapliga processen. Men utvecklingen inom t ex psykologi där förregistrering och *team science* snabbt har blivit vanligare tyder på att det går att ändra normer till det bättre (se t ex Nosek m fl 2018).

REFERENSER

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

Arrow, K J m fl (2008), ”The Promise of Prediction Markets”, *Science*, vol 320, s 877.

Benjamin, D J m fl (2017), ”Redefine Statistical Significance”, *Nature Human Behaviour*, vol 1, publicerad online.

Brodeur, A, M Lé, M Sangnier och Y Zylberberg (2016), ”Star Wars: The Empirics Strike Back”, *American Economic Journal: Applied Economics*, vol 8, s 1–32.

Button, K S m fl (2013), ”Power Failure: Why Small Sample Size Undermines the Reliability of Neuroscience”, *Nature Reviews Neuroscience*, vol 14, s 365–376.

Callaway, E (2011), ”Report Finds Massive Fraud at Dutch Universities”, *Nature*, vol 479, s 15.

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

Camerer, C F, A Dreber och M Johannesson (2018), ”Replication and other Practices for Improving Scientific Quality in Experimental Economics”, under utgivning i Schram, A och A Ule (red), *Handbook of Research Methods and Applications in Experimental Economics*.

Casey, K, R Glennerster och E Miguel (2012), ”Reshaping Institutions: Evidence on Aid Impacts Using a Pre-Analysis Plan”, *Quarterly Journal of Economics*, vol 127, s 1755–1812.

Chang, A och P Li (2015), ”Is Economics Research Replicable? Sixty Published Papers from Thirteen Journals Say ’Usually Not’”, Feds Working Paper 2015–083, Washington.

Cumming, G (2008), ”Replication and p Intervals: p Values Predict the Future only Vaguely, but Confidence Intervals Do Much Better”, *Psychological Science*, vol 3, s 286–300.

Dayenas, E m fl (1988), ”Human Basophil Degranulation Triggered by Very Dilute Antiserum against IgE”, *Nature*, vol 333, s 816–818.

Dewald, W G, J G Thursby och R G Andersson (1986), ”Replication in Empirical Economics: The Journal of Money, Credit and Banking Project”, *American Economic Review*, vol 76, s 587–603.

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 (2016), ”Many Labs 3: Evaluating Participant Pool Quality across the Academic Semester via Replication”, *Journal of Experimental Social Psychology*, vol 67, s 68–82.

Franco, A, N Malhotra och G Simonovits (2014), ”Publication Bias in the Social Sciences: Unlocking the File Drawer”, *Science*, vol 345, s 1502–1505.

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.

Gelman, A och H Stern (2006), ”The Difference between ’Significant’ and ’Not Significant’ is Not Itself Statistically Significant”, *American Statistician*, vol 60, s 328–331.

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

Hanson, R, (2007), ”Logarithmic Market Scoring Rules for Modular Combinatorial Information Aggregation”, *Journal of Prediction Markets*, vol 1, s 3–15.

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.

Johnson V E, R D Payne, T Wang, A Asher och M Soutrik (2017), ”On the Reproducibility of Psychological Science”, *Journal of the American Statistical Association*, vol 112, s 1–10.

Klein, R A m fl (2014), ”Investigating Variation in Replicability: A ’Many Labs’ Replication Project”, *Social Psychology*, vol 45, s 142–152.

Leamer, E (1983), ”Let’s Take the Con Out of Econometrics”, *American Economic Review*, vol 73, s 31–43.

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

McCullough, B D, K A McGeary och T D Harrison (2006), ”Lessons from the JMCB Archive”, *Journal of Money, Credit and Banking*, vol 38, s 1093–1107.

McCullough, B D och H D Vinod (2003), ”Verifying the Solution from a Nonlinear Solver: A Case Study”, *American Economic Review*, vol 93, s 873–892.

Miguel, E m fl (2014), ”Promoting Transparency in Social Science Research”, *Science*, vol 343, s 30–31.

Munafò, M R m fl (2017), ”A Manifesto for Reproducible Science”, *Nature Human Behavior*, vol 1, artikel 0021.

Nosek, B A m fl (2015), ”Promoting an Open Research Culture: Author Guidelines for Journals Could Help to Promote Transparency, Openness, and Reproducibility,” *Science*, vol 348, s 1422.

Nosek, B A, C R Ebersole, A DeHaven och D M Mellor (2018), ”The Preregistration Revolution”, under utgivning i *Proceedings for the National Academy of Sciences*.

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.

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.

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.

Verhagen, J och E-J Wagenmakers (2014), ”Bayesian Tests to Quantify the Result of a Replication Attempt”, *Journal of Experimental Psychology: General*, vol 143, s 1457–1475.

Wolfers, J och E Zitzewitz (2006), ”Interpreting Prediction Market Prices as Probabilities”, NBER Working Paper 12200.

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