Één onderzoek is géén onderzoek : het belang van replicaties voor de psychologische wetenschap

Recente fraudezaken trekken de robuustheid van wetenschappelijk onderzoek in twijfel. Toch is er een uitstekende manier om de betrouwbaarheid van onderzoek te garanderen: hetzelfde onderzoek meerdere keren uitvoeren. Waarom doen onderzoekers desondanks zelden replicatieonderzoek? Daniël Lakens c.s. leggen uit waarom een significante p-waarde betekenisloos is en alleen replicatieonderzoek de psychologische wetenschap kan redden.

AUTEURS: DANIËL LAKENS, ANTAL HAANS, SANDER KOOLE

HET BELANG VAN REPLICATIES VOOR DE PSYCHOLOGISCHE WETENSCHAP

ÉÉN ONDERZOEK IS GÉÉN ONDERZOEK

Psychologische wetenschappers proberen robuuste kennis te verwerven over het gedrag van mensen. De laatste jaren is die robuustheid regelmatig ter discussie gesteld. Maar al te vaak hoor je de opmerking: 'Wetenschap is ook maar een mening.' Spraakmakende fraudezaken zoals die van Diederik Stapel en Dirk Smeesters doen het vertrouwen in de wetenschap nog meer afnemen. Methodologen bekritiseren onderzoekers vanwege de grote flexibiliteit waarmee ze hun data analyseren. Onder psychologen heerst verontwaardiging over een artikel in een toptijdschrift waarin wordt beweerd dat mensen de toekomst kunnen voorspellen, zonder uit te leggen hoe mensen dit precies zouden moeten kunnen (Bem, 2011, zie ook LeBel & Peters, 2011; Wagenmakers, Wetzels, Borsboom & Van der Maas, 2011).

Deze ontwikkelingen motiveren onderzoekers om kritisch naar het eigen vakgebied te kijken. Er wordt dan ook op meerdere fronten gewerkt om de robuustheid van gepubliceerde onderzoeksresultaten te verbeteren. Een voorbeeld van zo'n initiatief is het *Open Science Framework* (opensciencefra mework.org). Binnen het Open Science Framework vindt een project plaats waarin psychologen over de hele wereld samenwerken aan het grootste replicatieonderzoek in de geschiedenis van de psychologie. Tientallen onderzoekteams voeren directe replicaties uit van bestaand onderzoek. Het doel van dit project is om het belang van replicatieonderzoek als een kwaliteitscontrole van wetenschappelijke bevindingen op de kaart te zetten en een voorlopige indicatie van de robuustheid van psychologische kennis te verkrijgen.

Bij replicatieonderzoek doet een wetenschapper een bestaand onderzoek zo goed mogelijk na, in de hoop hetzelfde resultaat te vinden. We lichten in dit artikel toe waarom replicatieonderzoek belangrijk is voor elke betrouwbare wetenschap, maar niettemin relatief weinig wordt uitgevoerd door wetenschappers. Daarna leggen we uit waarom nieuwe onderzoeksresultaten die bestaan uit één enkel onderzoek voorzichtig geïnterpreteerd moeten worden. We sluiten af met een voorzichtige optimistische blik op veranderingen binnen de wetenschappelijke cultuur.

Met dit artikel hopen we lezers duidelijk te maken dat het verstandig is de volgende vuistregel, gebaseerd op een oudhollands versje, in gedachten te houden: één onderzoek is

Methodologisch is het niet toegestaan extra proefpersonen te verzamelen als bij een eerste analyse blijkt dat het resultaat nét niet significant is

geen onderzoek, twee onderzoeken zijn een half onderzoek, drie onderzoeken zijn een betrouwbaar onderzoek.

EÉN ONDERZOEK IS GEEN ONDERZOEK Stel dat een nieuwe studie aantoont dat mensen die net naar een kinderliedje hebben geluisterd zich jonger voelen en zich kinderachtiger gedragen. Hoe groot is de kans dat andere onderzoeken dit effect ook vinden? Als een wetenschapper een gepubliceerd onderzoek herhaalt, zal het eerder beschreven effect soms niet gevonden worden. De replicatiepoging is dan niet succesvol. Het feit dat er onderzoeken in wetenschappelijke tijdschriften staan die niet succesvol gerepliceerd kunnen worden, is in principe onwenselijk. Tegelijkertijd is het een praktische werkelijkheid. Het is belangrijk om goed te begrijpen waarom sommige replicatiepogingen niet succesvol zullen zijn, vooral omdat er maar al te vaak een schijn van absoluutheid om wetenschappelijk onderzoek heen hangt.

Wetenschappers tolereren een bepaalde kans op fouten als ze conclusies trekken over hun bevindingen. Deze foutenmarge schept een balans tussen betrouwbaarheid en praktische haalbaarheid van wetenschappelijk onderzoek. In psychologisch onderzoek wordt deze foutenmarge meestal op vijf procent gesteld. Deze marge heeft een vergelijkbare rol als de maximumsnelheid in het verkeer. Als auto's niet harder zouden mogen rijden dan twintig kilometer per uur, voorkomen we weliswaar ongelukken maar we komen ook nergens meer. De meeste onderzoekers weten dat een statistische toets een significantie of *p*-waarde moet hebben die kleiner is dan .05. Deze *p*-waarde beantwoordt de volgende vraag: hoe waarschijnlijk is het waargenomen (of een nog extremer) onderzoeksresultaat – bijvoorbeeld dat mensen die een kinderliedje horen, zich kinderachtiger gedragen –, *wanneer we aannemen dat er geen effect bestaat in de volledige groep van mensen (de populatie) waarover men een uitspraak wil doen* – als je alle mensen op aarde een kinderliedje laat horen, gedraagt men zich niet anders dan normaal.

Zoals Cohen (1994) uitlegt hebben mensen vaak moeite met het interpreteren van dit soort *voorwaardelijke* kansen (de kans op X *gegeven de voorwaarde* dat Y nul is). Een veel gemaakte fout is het verwisselen van de kans op een gebeurtenis (de kans dat X gevonden wordt) en de voorwaarde (geen populatie-effect, of de voorwaarde dat Y nul is). Zo menen veel onderzoekers ten onrechte dat een *p*-waarde van .05 laat zien dat er 95% kans is dat er een effect in de populatie bestaat.

Een andere veelgemaakte denkfout is om op basis van een *p*-waarde van .05 te concluderen dat er een kans van 95% is dat een exacte replicatie van het onderzoek wéér een significant resultaat zal laten zien. In een steekproef van Oakes (1986) maakte zestig procent van de psychologische onderzoekers zich schuldig aan deze 'replicatiedenkfout'. In vervolgonderzoek van Haller en Krause (2002) dacht 49% van de onderzoekers verkeerd over de *p*-waarde. Het is daarom niet verwonderlijk dat onderzoekers de kans overschatten dat een onlangs voor het eerst gevonden onderzoeksuitkomst in een replicatieonderzoek wederom eenzelfde significant effect zal laten zien (Tversky & Kahneman, 1971).

Wat is dan wel de kans dat een gerapporteerd effect in de wetenschappelijke literatuur echt bestaat en geen toevallige bevinding is? Die vraag is moeilijk (en misschien zelfs onmogelijk) te beantwoorden op basis van één enkel onderzoek (Miller, 2009). Miller en Schwarz (2011, p. 359) concluderen daarom dat '*the initial result actually says hardly anything about what percentage of replication attempts should be*

successful.' Ioannidis (2005) legt in zijn artikel 'Why most published research findings are false' uit dat vooral kleinschalige onderzoeken naar zwakke (maar spannende of 'sexy') effecten kans maken om op toeval te berusten. Bij dit soort onderzoeken geeft statistische significantie bijna geen enkele garantie dat het effect echt is. Om te laten zien dat een dergelijk effect betrouwbaar is, moet het gerepliceerd worden (Cohen, 1994; Schmidt, 2009).

Dit is belangrijk om te beseffen voor wetenschappers die een enkel onderzoek publiceren, en ook voor journalisten die een artikel schrijven over dat enkele onderzoek en voor mensen die onderzoeksresultaten willen toepassen in de praktijk. Eén onderzoek is geen onderzoek. Een significant nieuw resultaat is aanleiding om een idee verder te onderzoeken – maar ook niets meer dan dat.

PUBLICATIEDRUK

Als het resultaat van een enkel onderzoek zo weinig zegt over de vraag of een nog niet eerder gevonden effect daadwerkelijk bestaat, waarom bestaan artikelen dan niet altijd uit een nieuw onderzoek en in ieder geval uit één directe replicatie van het nieuwe onderzoek? Er zijn diverse oorzaken te noemen, en de meeste van deze oorzaken zijn al lang bekend (bijv. Smith, 1970).

Eén reden is dat tijdschriften artikelen vaak kort en bondig willen houden. Soms heeft een onderzoeker wel meerdere onderzoeken gedaan maar vraagt het wetenschappelijke tijdschrift de auteurs om maar één onderzoek te rapporteren. De ruimte in papieren wetenschappelijke tijdschriften was van oudsher beperkt, en replicaties zijn volgens veel redacteuren de extra ruimte niet waard. Wie wil weten of wetenschappers hun onderzoek zelf al hebben gerepliceerd, raden wij dan ook aan contact op te nemen met de betreffende auteurs.

Een tweede reden dat er artikelen met maar één studie zijn, is dat een onderzoeker zelf niet meer tijd heeft of wil besteden aan een onderzoekslijn. Misschien is het onderzoek een kleine variant op bestaand onderzoek, wellicht laat het zien dat een laboratoriumonderzoek ook in het veld werkt. Als onderzoek direct en logisch voortbouwt op eerder robuust onderzoek, kan een wetenschapper een enkele studie betrouwbaar genoeg vinden. Het kan ook dat een onderzoeker zelf geen tijd meer heeft om opnieuw onderzoek naar de hypothese te doen en probeert andere wetenschappers daartoe te motiveren. De onderzoeker publiceert dan vooral een idee.

Waarschijnlijk de belangrijkste reden dat er artikelen gepubliceerd worden waarin maar één onderzoek beschreven wordt, is het feit dat het voor de carrière van een wetenschapper bevorderlijk is om zoveel mogelijk wetenschappelijke bevindingen te publiceren. Het maakt momenteel nog relatief weinig uit of je een artikel met één onderzoek publiceert of een artikel dat meerdere onderzoeken beschrijft. Als beginnende onderzoekers solliciteren naar een baan, of als gevorderde onderzoekers promotie willen maken, worden vaak simpelweg het aantal publicaties geturfd. Bij dit turfwerk telt het aantal onderzoeken per publicatie niet. Het publiceren van drie artikelen die ieder uit één enkel onderzoek bestaan, loont daardoor meer dan het publiceren van twee artikelen waarin je een effect twee keer repliceert. De wetenschappelijke kwaliteit van onderzoek komt door deze *publicatiedruk* in het gedrang.

Dat onderzoekers niet beloond worden om hun eigen werk te repliceren is eigenlijk vreemd en wetenschappelijk gezien problematisch. Dat replicatieonderzoeken bijna niet te publiceren zijn, versterkt dit probleem. Als iedereen af en toe onderzoek van een ander zou herhalen en deze replicatieonderzoeken kon publiceren in een wetenschappelijk tijdschrift, dan zou vanzelf duidelijk worden welke onderzoeken wel betrouwbaar zijn en welke niet. Helaas accepteren traditionele tijdschriften replicatieonderzoeken zelden voor publicatie. Zowel succesvolle als niet-succesvolle replicaties verdwijnen in de la van de onderzoeker en worden niet gedeeld met andere onderzoekers. Het belang van replicaties wordt daarmee in feite miskend. Binnen elke serieuze wetenschap zou innovatief onderzoek met onverwachte resultaten hand in hand moeten gaan met replicaties van bestaand onderzoek, toegepaste onderzoeken én theoretische integratie van onderzoeksresultaten. Alleen op die manier zal de psychologie robuuste en toepasbare kennis opleveren.

TWEE ONDERZOEKEN ZIJN EEN HALF ONDERZOEK

Soms doet men replicatieonderzoek af als het simpelweg overdoen van iets wat we al wisten. Replicaties leiden echter wel degelijk tot nieuwe inzichten. Ze laten zien of een effect betrouwbaar is of niet. De verklaring die onderzoekers voor het effect geven hoeft niet correct te zijn, maar het effect zelf is in ieder geval reproduceerbaar.

Sommige onderzoekers menen dat vooral de eerste exacte replicatie van een bestaand onderzoek een grote toename in betrouwbaarheid van het effect geeft (Tsang & Kwan, 1999).

Deze overtuiging heeft niet zozeer te maken met statistische overwegingen, maar met de mogelijke flexibiliteit waarvan onderzoekers gebruik kunnen maken als ze een eerste onderzoek analyseren en publiceren. Dirk Smeesters, de wegens fraude aan de Erasmus Universiteit Rotterdam ontslagen hoogleraar consumentengedrag, maakte waarschijnlijk in extreme mate gebruik van deze flexibiliteit (ook wel datamassage genoemd). Daardoor werden zijn onderzoeksresultaten onbetrouwbaar.

Simmons, Nelson & Simonsohn (2011) deden empirisch onderzoek naar de vraag of het luisteren naar het liedje 'When I'm Sixty-Four' van The Beatles proefpersonen ouder kon maken dan proefpersonen die tijdens het onderzoek naar een neutraal liedje luisterden. Hoewel deze vraagstelling natuurlijk onzinnig is (zelfs The Beatles kunnen de natuurwetten niet breken), vonden de auteurs dat de proefpersonen die naar dat liedje geluisterd hadden anderhalf jaar (en statistisch significant) ouder waren dan mensen in de controle conditie. Hoe is dit mogelijk? In hun artikel lieten ze zien hoe zij meerdere metingen hadden meegenomen maar vervolgens uitsluitend die metingen en toetsen rapporteerden die hun dwaze hypothese ondersteunden. De auteurs illustreerden hiermee op spectaculaire wijze dat bepaalde onderzoekspraktijken (waarvan sommige niet algemeen geaccepteerd worden, maar andere meer gebruikelijk zijn) ertoe kunnen leiden dat significantieniveaus in de praktijk veel flexibeler zijn dan de gangbare foutenmarge van vijf procent. Zo lijken onbetrouwbare resultaten statistisch significant. Afhankelijk van de vele keuzes die onderzoekers soms kunnen maken, ligt de daadwerkelijke foutenmarge al snel boven de tien procent en kan die zelfs oplopen tot boven de vijftig procent.

Het is zeker niet zo dat alle onderzoekers zo flexibel omgaan met hun data als Simmons en collega's (2011) doen voorkomen. Veel onderzoekers wijken niet af van het wetenschappelijke ideaal waarbij je van tevoren een specifieke onderzoeksvraag stelt en alleen data verzamelt en analyses doet die nodig zijn om die vooraf vastgestelde onderzoeksvraag te beantwoorden. Het is echter naïef te veronderstellen dat onderzoekers hun onderzoeksvraag nooit achteraf wat bijstellen of bepaalde analysetechnieken kiezen met de meest voordelige uitkomst voor de statistische toets (John, Loewenstein & Prelec, 2012).

Deze handelswijze wijkt af van het wetenschappelijke ideaal, maar is pragmatisch vaak in meer of mindere mate verdedigbaar. Zo is het methodologisch niet toegestaan extra proefpersonen te verzamelen als bij een eerste analyse blijkt dat het onderzoeksresultaat nét niet significant is. Correct zou zijn om op basis van de verzamelde data uit te rekenen hoeveel deelnemers je had moeten verzamelen en het hele onderzoek opnieuw te draaien met meer deelnemers. Onderzoekers kiezen er echter vaak voor om tijd en belastinggeld te besparen door alleen extra deelnemers te verzamelen en de oude en nieuwe data te combineren in de uiteindelijke analyse.

De flexibiliteit van onderzoekers bij het rapporteren van gegevens verhoogt het belang van exacte replicaties. Bij een exacte replicatie is er voor die flexibiliteit geen ruimte meer. Onderzoekers die een directe replicatie uitvoeren kunnen niet kiezen hoe ze hun data analyseren, zij moeten het eerdere onderzoek nauwkeurig nadoen. Feitelijk is deze replicatie het eerste onderzoek dat voldoet aan het wetenschappelijke ideaal en dat garandeert dat het onderzoek werkelijk een foutenmarge van vijf procent heeft. Een eerste succesvolle replicatie geeft om deze reden een grote toename in de betrouwbaarheid van een effect, zelfs al blijft het te vroeg om met grote zekerheid te zeggen dat het effect echt is. Daarom is de volgende vuistregel gerechtvaardigd: twee onderzoeken zijn een half onderzoek.

DRIE ONDERZOEKEN ZIJN EEN BETROUWBAAR ONDERZOEK

Met het laatste deel van het aangepaste kinderversje – drie onderzoeken zijn een betrouwbaar onderzoek – permitteren we ons enige dichterlijke vrijheid. Correcter zou zijn te stellen: veel onderzoeken zijn een betrouwbaar onderzoek. Het is immers net zo goed mogelijk dat een replicatieonderzoek geen significant effect laat zien terwijl dat effect er wel is, als dat een origineel onderzoek een significant effect laat zien dat er eigenlijk niet is. Net als mensen zien wetenschappers dus soms iets wél terwijl dat er niet is, en soms zien ze juist iets niet terwijl dat er wel is. Dat geldt zowel voor het eerste onderzoek als voor het replicatieonderzoek. Dit klinkt misschien raar voor mensen die gewend zijn dat de wetenschap onderzoeksresultaten opdeelt in twee groepen: significant (en dus 'echt') en niet significant (en dus 'niet echt'). Deze tweedeling is een hardnekkige illusie binnen de wetenschap. Om Rosnow en Rosenthal (1989, p. 1277) te citeren: *'Surely, God loves the .06 nearly as much as the .05.'*

Een *p*-waarde wekt slechts de schijn van absoluutheid. De werkelijkheid is dat de *p*-waarde van een onderzoek, zelfs als je een experiment perfect repliceert, keer op keer enorm kan verschillen. Zo simuleerde Cumming (2008) met een computerprogramma vijfentwintig keer hetzelfde

denkbeeldige onderzoek. Er was steeds gemiddeld genomen over alle virtuele proefpersonen een echt verschil tussen de experimentele groep en de controlegroep, maar de score van elke virtuele proefpersoon varieerde (net zoals in echte experimenten). De simulatie trok willekeurig een aantal proefpersonen uit de virtuele populatie. Onder de gebruikelijke omstandigheden in psychologisch onderzoek varieerden de *p*-waardes van de vijfentwintig gesimuleerde onderzoeken van $p < .001$ tot $p = .759$.

Deze simulatie illustreert dat niet elk experiment een statistisch significant verschil oplevert, zelfs als er een echt verschil bestaat in de populatie. Of een onderzoek een significant effect oplevert als dat er ook echt is, hangt af van de grootte van het effect, het aantal deelnemers en de statistische foutenmarge. Het gemiddelde statistische onderscheidingsvermogen (*power*) in psychologisch onderzoek wordt geschat op vijftig procent (Schmidt & Hunter, 1997). Als er echt een significant effect is in de populatie, dan is de kans dat een gemiddeld onderzoek dit effect bevestigt dus even groot als de uitkomst van kop bij het opgooien van een munt. Dat is niet hoog, en veel onderzoekers maken zich terecht zorgen over de mogelijkheid dat onderzoekers geen effecten vinden terwijl die er wel zijn.

Een hoger statistisch onderscheidingsvermogen kan worden verkregen met grotere aantallen proefpersonen (honderden, maar soms tienduizenden mensen). Dit is bij onderzoek met vragenlijsten in principe nog haalbaar, en ook het internet biedt steeds meer mogelijkheden voor onderzoek met grote groepen deelnemers (zie bijvoorbeeld Project Implicit op http://implicit.harvard.edu). Toch blijven zelfs de mogelijkheden van internet beperkt en heerst onder psychologen de opvatting dat het praktisch onhaalbaar is om elk empirisch onderzoek met zulke grote aantallen proefpersonen te doen. Hoewel mensen in replicaties vaak streven naar de hoogst haalbare power, wordt een power van .80 acceptabel gevonden (Cohen, 1990). Daarom zal niet elk replicatieonderzoek een significant effect geven, zelfs als dit er wel is.

Als beginnende onderzoekers solliciteren naar een baan worden simpelweg hun aantallen publicaties geturfd

PSYCHFILEDRAWER.ORG

De moraal van dit verhaal is dat één onderzoek weinig zegt over het al dan niet bestaan van een effect. Twee onderzoeken met een significant effect is al veel minder waarschijnlijk. Maar je weet pas echt zeker dat een onderzoeksresultaat betrouwbaar is, als je het meerdere keren kunt repliceren. Eén onderzoek is geen onderzoek, twee onderzoeken zijn

een half onderzoek, drie (en bij voorkeur meer) onderzoeken zijn een betrouwbaar onderzoek. Door in een meta-analyse naar de effecten over verschillende onderzoeken te kijken, is met toenemende betrouwbaarheid vast te stellen of een effect echt bestaat, en misschien nog wel belangrijker, hoe groot het effect ongeveer is.

Dat het effect betrouwbaar is en voor waar aangenomen kan worden, is belangrijk voor wetenschappers die theorieën over menselijk gedrag ontwikkelen. Deze theorieën moeten immers niet gebaseerd zijn op conclusies van onbetrouwbare onderzoeken. Hoe groot het effect is, heeft vooral praktisch nut. Bij het toepassen van onderzoeksresultaten is het van belang om te weten welke manipulatie van de beschikbare alternatieven het grootste effect laat zien, bijvoorbeeld bij het behandelen van patiënten. Het zou daarom goed zijn als replicatieonderzoek relatief meer beloond werd. De Nederlandse Organisatie voor Wetenschappelijk Onderzoek (NWO) zou bijvoorbeeld bij het uitdelen van subsidie minder kunnen kijken naar aantallen publicaties en meer naar hoe vaak het werk van een onderzoeker succesvol is gerepliceerd. Dit helpt om de psychologische wetenschap robuust en toepasbaar te maken.

Veel wetenschappers zien als bezwaar van meer replicatieonderzoek dat er minder nieuwe kennis wordt verzameld. Ze geven er de voorkeur aan om conceptuele replicaties uit te voeren, waarbij ze niet precies hetzelfde nogmaals doen, maar het onderzoek met andere manipulaties en/of andere maten nogmaals uitvoeren. Een succesvolle conceptuele replicatie is waardevol, omdat het laat zien dat een effect te generaliseren is naar andere methoden en omstandigheden. Een niet succesvolle conceptuele replicatie geeft echter geen enkel uitsluitsel over de vraag of het originele effect betrouwbaar is of niet. Zoals LeBel en Peters (2011) opmerken kan een onderzoeker zo doorgaan tot er van een groot aantal onderzoeken een paar varianten een goed resultaat hebben laten zien, terwijl onderzoeken die niet hebben gewerkt worden genegeerd. Een artikel lijkt dan een mooi verhaal te vertellen, maar van elk onderzoek is het nog steeds onzeker of het effect wel echt bestaat. Een artikel met meerdere conceptuele replicaties heeft meer potentie dan een 'één-onderzoek artikel', maar de toekomst moet uitwijzen hoe betrouwbaar de resultaten zijn. Er is iets voor te zeggen dat niet de toekomst maar de onderzoeker zelf laat zien hoe betrouwbaar de resultaten zijn, gevolgd door replicaties door collega-onderzoekers van andere universiteiten.

Natuurlijk doen onderzoekers niet alleen maar nieuw onderzoek. Een veelgebruikte strategie is een replicatie met uitbreiding. Hierbij repliceert een onderzoeker eerder onderzoek, maar wordt ook een nieuwe experimentele conditie toegevoegd waaruit nieuwe inzichten zijn af te leiden. Deze benadering heeft belangrijke voordelen. Omdat replicatieonderzoek momenteel weinig gepubliceerd wordt, weten onderzoekers vaak niet hoe betrouwbaar de resultaten zijn waarop ze verder willen bouwen. Door een onderzoek te doen dat bestaat uit een replicatie met uitbreiding kun je zowel een nieuw idee testen als controleren of het bestaande onderzoek werkt.

Een p-waarde wekt slechts de schijn van absoluutheid

Het zou natuurlijk nog transparanter zijn als succesvolle en niet-succesvolle replicaties van bestaand onderzoek voor iedereen toegankelijk zouden zijn. Dit is het idee achter PsychFileDrawer.org, een website waarop onderzoekers de uitkomsten van hun replicatieonderzoek kunnen delen. Gegeven het belang van replicatieonderzoek voor empirische wetenschappen is het vreemd dat er eerst een aparte internetsite moest komen om gelukte of niet gelukte replicaties te kunnen delen met andere onderzoekers. Replicaties blijven op deze manier ook vaak nog weinig zichtbaar, omdat veel onderzoekers (en media) de weg nog niet weten te vinden naar PsychFileDrawer.org. Een dergelijke website doet dus eigenlijk nog steeds te kort aan het immense wetenschappelijke belang van replicaties.

Replicaties zijn onmisbaar voor elke empirische wetenschap. Alle goed uitgevoerde replicatieonderzoeken zouden daarom gewoon net zo toegankelijk moeten zijn als nieuwe onderzoeken. Als wetenschappelijke conclusies voortaan gebaseerd worden op meta-analyses over drie (of meer) onderzoeken, dan is het noodzakelijk dat deze meta-analyses gedaan worden over zowel succesvolle als niet-succesvolle replicaties. Alleen dan zijn de conclusies over of

een effect wel of niet bestaat, en over hoe groot het is, betrouwbaar.

PSYCHOLOGIE ALS SLOW SCIENCE Soms wordt er over wetenschappelijk onderzoek gesproken alsof het absolute waarheden oplevert. Andere keren zijn mensen geneigd te zeggen dat wetenschap ook maar een mening is. Beide uitgangspunten zijn niet correct. Uitkomsten van wetenschappelijk onderzoek zijn niet zwart-wit. Cumulatieve wetenschap geeft met toenemende waarschijnlijkheid aan of effecten bestaan of niet. Let wel: dit geldt alleen voor cumulatieve wetenschap. Een toename tot een aan zekerheid grenzende waarschijnlijkheid is alleen mogelijk als effecten meerdere keren onderzocht worden. Wetenschap is nooit slechts een mening. Wetenschappelijke conclusies worden immers altijd, ook als er maar één onderzoek gedaan is, beperkt door de verzamelde data. Maar omdat van een nieuw wetenschappelijk onderzoek meestal nog zo onduidelijk is of het erin gestelde effect bestaat of niet (vooral bij kleine onderzoeken met kleine effecten), mogen we van rationeel denkende mensen verwachten dat ze het onderzoeksresultaat nog niet serieus nemen.

Stel, je wilt met de trein naar huis. De trein die je normaal neemt, heeft een onbekende vertraging. Je kunt ook omrijden via een andere route. Via sociale media vraag je wat je moet doen en 28 mensen reageren. De helft koos in het verleden een alternatieve route. Statistisch zijn die mensen een heel klein beetje, maar precies significant (de *p*-waarde is .049), sneller op de bestemming. Wat moet er op basis van jouw onderzoek door de omroepinstallatie geroepen worden?

De huidige situatie is maar al te vaak dat de omroep-installatie (lees: de media, maar ook de wetenschap zelf) roept dat je sneller thuis bent als je de alternatieve route neemt. Dat is evenwel misleidend. De omroepinstallatie zou moeten zeggen dat het verschil tussen de normale en alternatieve route met 95% zekerheid ligt tussen 45 minuten sneller en 35 minuten trager. De kans dat de reizigers op het perron gemiddeld sneller thuis zullen zijn (op basis van deze kleine steekproef) is niet te voorspellen. Maar als je er optimistisch naar kijkt, zou het ongeveer vijftig procent kunnen zijn. Of de reizigers het zelf, op basis van deze waarschijnlijkheden, de moeite waard vinden om uit te zoeken wat de alternatieve route is en naar een ander spoor te lopen, is hun eigen keuze. Maar we kunnen het reizigers moeilijk kwalijk nemen als ze zeggen: 'Eén onderzoek is geen onderzoek. Ik wacht nog even voordat ik mijn acties verander, want drie onderzoeken zijn pas een betrouwbaar onderzoek.'

Voor psychologen kan het publiceren van een eerste onderzoeksresultaat wel waardevol zijn. Andere onderzoekers kunnen zich ook bezig gaan houden met dat effect, zodat sneller duidelijk wordt of het echt bestaat of niet. Pas als een effect een aantal keer gerepliceerd is, en de waarschijnlijkheid dat het echt bestaat groot genoeg is, wordt het relevant voor een breder publiek dat onderzoeks-resultaten kan toepassen. Maar zelfs als effecten succesvol gerepliceerd zijn, kan de grootte van het gevonden effect zo klein zijn dat het misschien theoretisch belangrijk is maar tegelijk praktisch irrelevant.

Mitchell (2012) vergeleek de effecten van veldstudies en labstudies in 82 meta-analyses. Waar de effecten die organisatiepsychologen bestuderen goed te repliceren zijn in de praktijk, bleken veel effecten die sociaal psychologen onderzoeken minder duidelijk aanwezig in de veldstudies. Een mogelijke verklaring hiervoor is dat de effecten in de meta-analyses binnen de organisatiepsychologie veel groter zijn dan die in de sociale psychologie, en kleine effecten in het lab worden nu eenmaal makkelijker verstoord in het minder controleerbare echte leven.

De stand van zaken wat betreft replicatieonderzoek is niet anders in psychologisch onderzoek dan in veel andere wetenschappelijke disciplines. Maar het onderzoeksklimaat lijkt te veranderen. Onderzoekers spreken zich uit voor *slow science* waarin, net zoals bij de *slow food*-beweging, kwaliteit belangrijker wordt gevonden dan kwantiteit (Alleva, 2006). Dit thema leeft ook binnen de psychologie (bijvoorbeeld Mummendey, 2012; Van den Bos, 2011). Slow science richt zich op het belang van theorievorming, replicatie en toepassing. Niet toevalligerwijs de drie aspecten van de wetenschappelijke cyclus die momenteel minder aandacht krijgen dan het publiceren van veel onderzoek dat vernieuwende voorspellingen toetst en onverwachte resultaten rapporteert.

OPEN SCIENCE FRAMEWORK

Het in de introductie besproken Open Science Framework probeert actief bij te dragen aan een robuuste sociale psychologie door het belang van replicatie te benadrukken en door een eerste stap te zetten in het actief uitvoeren van replicatieonderzoeken. Het samenwerkingsverband binnen het Open Science Framework heeft duidelijke procedures ontwikkeld om replicatieonderzoek gedegen uit te voeren, en probeert op verschillende manieren te stimuleren dat

replicatieonderzoek een integraal onderdeel van de wetenschappelijke cyclus wordt. Tegelijk is het een prachtig voorbeeld van hoe wetenschappers samen kunnen werken om verbeteringen te bewerkstelligen. Al deze onderzoekers steken hun tijd en energie in onderzoek dat voor elk individu op zich niet relevant is, maar voor de psychologische wetenschap als discipline zeer waardevol kan blijken te zijn. Nu onderzoeken deze wetenschappers de repliceerbaarheid van psychologisch onderzoek. Maar stelt u zich eens voor dat deze of soortgelijke groepen onderzoekers, zonder te letten op het aantal publicaties dat ze individueel scoren, zich gezamenlijk storten op maatschappelijk relevante vragen en robuuste kennis verzamelen die toegepast kan worden om het leven van mensen te verbeteren...

Het zijn spannende tijden voor psychologisch onderzoekers.

OVER DE AUTEURS

Dr. D. Lakens en dr. A. Haans zijn als universitair docent verbonden aan de Human Technology Interaction Group van de Technische Universiteit Eindhoven, Postbus 513, 5600 MB Eindhoven. Dr. S. Koole is verbonden aan de Faculteit der Psychologie en Pedagogiek, afdeling Klinische Psychologie van de Vrije Universiteit Amsterdam. Voor correspondentie over dit artikel: D.Lakens@tue.nl.

Summary

One study is no study: The importance of replication for psychological science

D. Lakens, A. Haans, S.L. Koole

Recent criticisms on the way psychologists analyze their data, as well as cases of scientific fraud, have led both researchers and the general public to question the reliability of psychological research. At the same time, researchers have an excellent tool at their disposal to guarantee the robustness of scientific findings: replication studies. Why do researchers rarely perform replication studies? We explain why *p*-values for single studies fail to provide any indication of whether observed effects are real or not. Only cumulative science, where important effects are demonstrated repeatedly, is able to address the challenge to guarantee the reliability of psychological findings. We highlight some novel initiatives, such as the Open Science Framework, that aim to underline the importance of replication studies.

Literatuur

Alleva, L. (2006). Taking time to savour the rewards of slow science. *Nature*, *443*, 271. doi:10.1038/443271e.

Bem, D. J. (2011). Feeling the future: Experimental evidence for anomalous retroactive influences on cognition and affect. *Journal of Personality and Social Psychology*, *100*, 407–425. doi:10.1037/a0021524.

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.

Cumming, G. (2008). Replication and *p* intervals: *p* values predict the future only vaguely, but confidence intervals do much better. *Perspectives on Psychological Science*, *3*, 286–300.

Haller, H. & Krauss, S. (2002). Misinterpretations of significance: A problem students share with their teachers? *Methods of Psychological Research*, *7*, 1–20.

Ioannidis J. P. (2005). Why most published research findings are false. *PLoS medicine*. 2: e124. doi:10.1371/journal.pmed.0020124.

John, L., Loewenstein, G. & Prelec, D. (2012). Measuring the prevalence of questionable research practices with incentives for truth-telling. *Psychological Science*, *23*, 524-532. doi: 10.1177/0956797611430953.

LeBel, E. P. & Peters, K. R. (2011). Fearing the future of empirical psychology: Bem's (2011) evidence of psi as a case study of deficiencies in modal research practice. *Review of General Psychology*, *15*, 371-379. doi: 10.1037/a0025172.

Mitchell, G. (2012). Revisiting truth or triviality: The external validity of research in the psychological laboratory. *Perspectives on Psychological Science*, *7*, 109-117. doi: 10.1177/1745691611432343.

Miller, J. (2009). What is the probability of replicating a statistically significant effect? *Psychonomic Bulletin & Review*, *16*, 617–640. doi:10.3758/PBR.16.4.617.

Miller, J. & Schwarz, W. (2011). Aggregate and individual replication probability within an explicit model of the research process. *Psychological Methods*, *16*, 337–360.

Mummendey, A. (2012) Scientific Misconduct in Social Psychology – Towards a Currency Reform in Science. *European Bulletin of Social Psychology*, *24*, 4-7.

Oakes, M. (1986). *Statistical inference: A commentary for the social and behavioral sciences*. New York: Wiley.

Rosnow, R.L. & Rosenthal, R. (1989). Statistical procedures and the justification of knowledge in psychological science. *American Psychologist*, *44*, 1276–1284.

Schmidt, S. (2009). Shall we really do it again? The powerful concept of replication is neglected in the social sciences. *Review of General Psychology*, *13*, 90–100. doi:10.1037/a0015108.

Simmons, J. P., Nelson, L. D. & Simonsohn, U. (2011). False- positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science*, *22*, 1359–1366.

Smith jr., N. C. (1970). Replication studies: A neglected aspect of psychological research. *American Psychologist*, *25*, 970-975.

Tsang, E., & Kwan, K. 1999. Replication and theory development in organizational science: A critical realist perspective. *Academy of Management Review*, *24*: 759–780.

Tversky, A., & Kahneman, D. (1971). Belief in the law of small numbers. *Psychological Bulletin*, *92*, 105–110.

Van den Bos, K. (2011). We moeten af van de lijstjescultuur. Retrieved May 22, 2012, from http://www.dub.uu.nl/artikel/we-moeten-af-lijstjescultuur.html.

Wagenmakers, E.J., Wetzels, R., Borsboom, D. & Van der Maas, H. L. J. (2011). Why psychologists must change the way they analyze their data: The case of psi: Comment on Bem (2011). *Journal of Personality and Social Psychology*, *100*, 426–432. doi:10.1037/a0022790.