

BETROUWBAAR



WETENSCHAPPELIJK HERHAALONDERZOEK



Barsten in

Fouten in wetenschappelijk onderzoek komen vroeg of laat aan het licht omdat onderzoekers elkaars experimenten herhalen. Of dat denken we tenminste. Het dubbelchecken van andermans studies is echter allesbehalve populair. Er is een replicatiecrisis, klinkt het onder wetenschappers.

Reinout Verbeke / Illustraties: Han Van de Ven

Denken aan een professor doet je hoger scoren op een intelligentietest dan denken aan een hooligan. Het is een vinding van Nederlands sociaalpsycholoog Ap Dijksterhuis uit 1998, die het fenomeen *social priming* op de voorgrond bracht: ogenschijnlijk banale acties of gebeurtenissen sturen onze beslissingen. Vorig jaar probeerden vakgenoten van Dijksterhuis in negen onafhankelijke experimenten hetzelfde effect opnieuw te verkrijgen in een zogenoemd herhaalonderzoek of een replicatie. Dat lukte in geen van de negen proeven. Als het effect zou bestaan, dan zou je verwachten dat de meerderheid van de herhaalproeven bevestigt dat je het met een hoogleraar in gedachten beter doet op een IQ-test. Er

kwam een discussie op gang: had Dijksterhuis, die zijn precieze onderzoeksmethode niet wilde vrijgeven, gemorrel met de data? Voerden zijn vakgenoten de experimenten niet goed uit? Of spelen nog onbekende factoren een rol? De vragen zijn vandaag nog niet beantwoord. Twijfels bij veelgeciteerd onderzoek blijft niet beperkt tot psychologie. Twee jaar geleden probeerden 100 wetenschappers van het farmabedrijf Amgen 53 studies te repliceren die ze als mijlpalen beschouwden in het fundamenteel onderzoek naar kanker (veelbelovend basisonderzoek op weefsel of op proefdieren). Ze werkten vaak nauw samen met de onderzoekers van de oorspronkelijke studies om zeker te zijn dat de gebruikte tech-

betrouwbaarheid

nieken overeenkwamen. Verontrustend: slechts 6 van de 53 studies konden worden bevestigd. 'Dit zijn de studies waar farmabedrijven op vertrouwen om naar nieuwe medicijnen te zoeken', benadrukte Amgen-onderzoeksleider Glenn Begley het motief voor zijn initiatief. Het Duitse farmabedrijf Bayer Health Care legde in 2011 niet veel betere cijfers voor: in driekwart van de 67 studies die ze opnieuw deden, werden de bevindingen niet bevestigd.

Nieuwzucht

De schaarse pogingen om onderzoek systematisch te herhalen, maken duidelijk dat we veelbelovende studies niet zomaar mogen geloven. Waarom blijkt veel onderzoek niet solide? Flagante fraude – het verzinnen of manipuleren van data – is waarschijnlijk maar een klein deel van het antwoord. In 2005 bond John Ioannidis, epidemioloog van Stanford University, de kat de bel aan in het vakblad *PLOS Medicine*: hij onderzocht de belangrijkste oorzaken van zogenoemde 'valspositieve' onderzoeksuitkomsten, resultaten die alleen door toeval goed uitdraaien (een verband is 'statistisch niet significant' als er meer dan vijf procent kans is dat het door toeval voorkomt). Dat krijg je volgens zijn simulaties va-

ker bij te kleine steekproeven – te weinig proefpersonen bijvoorbeeld. De kans dat een vinding waar is, is volgens Ioannidis' studie ook kleiner als véél verbanden tegelijkertijd getest worden (tussen genen en ziektes bijvoorbeeld) en er te weinig voorselectie is van welk verband precies zal worden onderzocht. De kans op waarheid slinkt ook bij onderzoeksprocessen die 'kneedbaar' zijn

Er valt geen geld of eer te behalen met replicatieonderzoek

(waarbij het ontwerp of de methode onderweg aangepast kan worden), en naarmate de financiële belangen groter zijn. Hij concludeerde dat de meerderheid van de onderzoeksresultaten daarom fout is. Een van de redenen waarom de vele valspositieve uitkomsten niet worden ontdekt – en dus in de literatuur geciteerd blijven worden – is het gebrek aan herhaalonderzoek. Onderzoeken bevestigen of verwerpen wordt aangezien als de heilige graal van wetenschap. Zo riep Nobelprijswinnaar Daniel Kahneman van Princeton University



in de discussie rond Ap Dijksterhuis alle *priming*-onderzoekers op om een eindeloze ketting van herhaalonderzoek in gang te zetten. Alleen zo kon volgens hem ‘de troep’ collectief worden opgeruimd. En dat lijkt ook broodnodig: in de psychologie is slechts één procent van alle gepubliceerde onderzoek een replicatie. Goeie schattingen uit andere vakgebieden zijn er voorlopig niet, maar ook daar spreken we-

Explosie van papers

Door wetenschappelijke groeilanden als China en door de publicatiedruk is de hoeveelheid wetenschappelijke papers de voorbije decennia explosief toegenomen: elke twaalf jaar verdubbelt het aantal artikelen. In 2012 kreeg de medische databank PubMed er meer dan 1,1 miljoen artikelen bij, dat is liefst 126 per uur. Het totale aantal papers in 2012, in alle vakdomeinen, wordt op dubbel zoveel geschat. Het aantal klassieke vaktijdschriften bedraagt vandaag volgens database Scopus al meer dan 21.000. Vooral de online ‘*open access*’-tijdschriften nemen spectaculair toe: van 5.000 in 1997 tot 9.708 in 2014. De controle en de *peer review* bij veel (kleinere) tijdschriften is vaak ondermaats of afwezig, blijkt uit steekproeven met nepartikelen.

tenschappers over een replicatiecrisis, een nijpend tekort aan herhaalonderzoek.

‘Het is saai natuurlijk, en je moet wel heel gedreven zijn om godganse dagen andermans papers door de mangel te halen’, zegt methodoloog Michèle Nuijten van Tilburg University. ‘Ik ken ook geen enkele wetenschapper die carrière heeft gemaakt met herhaalonderzoek.’ Precies daar wringt het schoentje: het wetenschappelijke bedrijf beloont geen ‘trage wetenschap’. Bij fondsaanvragen – vandaag cruciaal om als onderzoeker aan de bak te blijven – tellen vooral je publicaties met spannende nieuwe bevindingen, en je krijgt herhaalonderzoek aan bijna geen enkel hoogstaand vaktijdschrift verkocht, omdat het ‘geen nieuws’ is. ‘Herhaalonderzoek krijgt bij ons weinig kans’, geeft Sabine Kleinert toe. Zij is Senior Executive Editor van het gereputeerde medische vakblad *The Lancet*. ‘Er zijn genoeg andere tijdschriften die replicaties wel publiceren.’ Maar als onderzoekers voor hun carrière in de toptijdschriften willen staan – en dat willen ze allemaal – zijn

Slechts 6 van de 53 belangrijke kankerstudies konden worden bevestigd

ze weinig bereid tijd en geld te spenderen aan herhaalonderzoek.

Randy Schekman, Nobelprijswinnaar Geneeskunde van vorig jaar, heeft alvast de knoop doorgehakt: zijn lab zal geen papers meer insturen naar de ‘luxetijdschriften’ *Cell*, *Nature* en *Science*. De publicatiedruk die ervan uitgaat zet ‘onderzoekers aan resultaten te masseren en alleen trendy onderzoek te doen, in plaats van belangrijk werk’.

Encore une fois

‘Het probleem wordt al vijftig jaar erkend, maar vandaag werken wetenschappers ook aan oplossingen’, zegt Nuijten. ‘Er is een aantal interessante replicatieprojecten aan de gang.’ Brian Nosek, hoogleraar psychologie aan University of Virginia, is initiatiefnemer van The Reproducibility Project. Meer dan 150 onderzoekers over de hele wereld herhalen studies uit 2008 die in drie prominente psychologietijdschriften verschenen, waaronder een studie van de Universiteit Gent. Met het project willen Nosek en co de betrouwbaarheid van het vakgebied testen, dat door de fraudezaak van Nederlands sociaalpsycholoog Diederik Stapel – 56 teruggetrokken studies, de meeste wegens verzonnen data – een flinke knauw

kreeg. ‘We doen dit omdat we van wetenschap houden’, zegt Nosek in *Science*. ‘Het doel is om de basiswaarden van wetenschap – transparantie, data delen, zelfkritiek en reproduceerbaarheid – in overeenstemming te brengen met de praktijk.’ In november vorig jaar kwamen 36 onderzoeksgroepen al collectief naar buiten met herhalingen van dertien experimenten over wereldwijd geaccepteerde theorieën in de psychologie: tien daarvan werden bevestigd. Eind mei bundelde het vaktijdschrift *Social Psychology* replicaties van nog eens veertien bekende psychologiestudies. In meer dan de helft van de gevallen was de herhaling geheel of gedeeltelijk mislukt. Onder meer bij experimenten die zouden aantonen dat een streep muziek invloed heeft op je keuzes, of dat bijgeloof je beter doet golfen.

Ook medische wetenschappers willen schoon schip maken. Science Exchange biedt medici de mogelijkheid hun studies tegen betaling te laten herhalen door een onafhankelijk team onderzoekers. Als een replicatie de resultaten van de oorspronkelijke studie bevestigt, krijgt het onderzoek een keurmerk van betrouwbaarheid. ‘Maar één herhaalpoging is nooit genoeg om het definitieve oordeel over een studie te vellen’, nuanceert Nuijten. ‘De kans op toevalstreffers blijft bestaan.’ Science Exchange kreeg subsidie om gedurende één jaar de vijftig meest geciteerde kankerstudies tussen 2010 en 2012 te repliceren. Op die manier zal duidelijk worden of de alarmerende cijfers van Amgen en Bayer steekhouden. Het verschil met de experimenten van de twee farmabedrijven is dat alle data en de methoden van het ‘Reproducibility Initiative’ volledig transparant zullen zijn.

‘Wij denken dat er, op lange termijn, een verschuiving zal komen: van het belonen van spectaculaire, onverwachte resultaten naar herhaalbare, kwalitatieve resultaten, die echt waar zijn’, zegt Elisabeth Iorns, voormalig kankeronderzoeker en CEO van Science Exchange. Voor zo’n verschuiving moet het beloningssysteem in de wetenschap veranderen, en dat zal nog heel wat discussie vergen. Sommige wetenschappers willen fondsverstrekkers ertoe brengen een vast deel van de budgetten opzij te zetten voor herhaalonderzoek. Vaktijdschriften zouden publicatieruimte moeten creëren voor replicaties én voor negatieve onderzoeksuitkomsten. Onder meer *Journal of Experimental Social Psychology* en *Journal of Negative Results* doen dat al, en op de website *Psych File Drawer* (psychfiledrawer.org) kun je herhaalpogingen uploaden. Een tiental vaktijdschriften, zoals *Cortex* en *F1000Research*, experimenteren

zelfs met een andere vorm van publiceren: een onderzoeker krijgt al toestemming voor publicatie op basis van de methode en het opzet van het onderzoek. Pas daarna wordt het experiment uitgevoerd. Met zulke initiatieven

Een verzonnen studie vol fouten raakte gepubliceerd in 157 tijdschriften

krijgen negatieve resultaten en herhaalonderzoek de kans om het daglicht te zien. Ook de broodheren van de wetenschap – onderzoeksinstituten en fondsverstrekkers – zouden herhaalonderzoek moeten waarderen in het carrièreparcours van hun wetenschappers, want nu ontbreekt elke stimulans.

Blinde vakgenoten

Het wetenschappelijke systeem heeft behalve herhaalonderzoek nog een andere belangrijke toetssteen: *peer review*. Een redacteur van een vaktijdschrift stelt enkele onafhankelijke experts aan om een ingezonden paper te controleren. Zij kunnen de paper goedkeuren, of grote en kleine herzieningen voorstellen. Die controle geldt als de ‘gouden standaard’ voor betrouwbaarheid. Maar gepubliceerde studies die worden teruggetrokken wegens fouten of fraude – dat waren er zo’n vierhonderd in 2013 – maken duidelijk dat de filter niet altijd goed werkt. En misschien zijn de mazen wel veel groter dan

China’s publicatiedruk

De wetenschappelijke output van China is de voorbije jaren sterk gestegen. Het aandeel Chinese artikelen in vakbladen – bijgehouden door de Science Citation Index – steeg tussen 2001 en 2011 met 9,5 procent. Daarmee staat het land op plaats twee, na de Verenigde Staten. China haalt in 2013 de vijfde plaats als het over citaties gaat – een graadmeter voor de invloed van die artikelen. Maar recent kwam aan het licht dat sommige onderzoeksinstituten de publicatiedruk op de spits drijven: je kunt er als onderzoeker je inkomen (gemiddeld 39.850 yuan of 4.628 euro per jaar) vertienvoudigen met één toppublicatie. Sommige onderzoeksinstituten hanteren de sleutel: 10.000 yuan maal 1 plus de impactfactor van het vaktijdschrift (dat is het gemiddeld aantal keren dat artikelen in dat tijdschrift worden geciteerd). Dus: staat je paper in *The Lancet*, met impactfactor 39,06, dan reiken sommige universiteiten daar een beloning van 400.600 yuan (46.526 euro) voor uit.

Privébedrijfjes maken misbruik van die torenhoge publicatiedruk, zo onthulde journalist Mara Hvistendahl eind vorig jaar in *Science*. Ze bieden kant-en-klare (verzonnen) papers aan, of regelen een publicatie in (nepversies van) bestaande vaktijdschriften. In China worden dus niet alleen dvd’s en kleren nagemaakt.

▲ Met het computer-programma SCIgen maak je met een muisklik een paper vol nonsens.

we denken. In oktober vorig jaar wist een verzonnen onderzoeker van een verzonnen instituut een verzonnen studie vol fouten in liefst 157 van de 304 aangeschreven wetenschappelijke vakbladen te publiceren. Slechts 60 procent van de geteste vaktijdschriften gebruikten überhaupt *peer review*. John Bohannon, de man achter de valstrik en redacteur van *Science*, had alleen de online *open access*-tijdschriften uitgedaagd, die gratis zijn en die hun inkomsten halen uit het geld dat onderzoekers betalen om te worden gepubliceerd (*Journal of International Medical Research* vroeg Bohannon zonder ook maar één kritische opmerking over zijn valse artikel 2.250 dollar). Maar in februari dit jaar hadden de klassieke, ‘gesloten’ vaktijdschriften ook prijs: twee grote uitgeversbedrijven moeten meer dan 120 papers terugtrekken die door het computerprogramma SCIgen werden gegenereerd en vol nonsens staan. Iedereen kan het online programma

ten willen herhalen ondervinden ook moeilijkheden om aan de oorspronkelijke methodes of de oorspronkelijk data te komen. In een recente enquête geeft slechts 45 procent van de onderzoekers aan bereid te zijn hun data te delen als een wetenschapper erom vraagt. Steeds meer vaktijdschriften en fondsverstreckers eisen nu dat wetenschappers al hun data centraal opslaan. En *Nature* biedt nu ook als eerste een checklist van achttien punten om te verzekeren dat alle technische en statistische informatie die cruciaal is voor het herhaalexperiment, gepubliceerd worden. Er is ook *publication bias*: vooral de statistisch significante, positieve resultaten worden gepubliceerd in vaktijdschriften, omdat daar nieuws in zit. Dat zorgt voor een tsunami aan positieve studies, in vakdomeinen zoals psychologie bevestigt zelfs tot 90 procent van de gepubliceerde studies de hypothese. Dat is te mooi om waar te zijn. ‘Maar die *publication bias* geldt vermoede-

dat wetenschappers van het Massachusetts Institute of Technology in 2005 maakten, gebruiken. Met één muisklik maakt het at random een paper voor je, inclusief tabellen, bijschriften, en referenties. Tussen 2008 en 2013 raakten zestien van zulke ‘artikelen’ in tijdschriften van de Duitse uitgeverij Springer en honderd in tijdschriften van het Amerikaanse Institute of Electrical and Electronic Engineers (IEEE). Zijn vakgenoten blind? ‘Nee, maar *peer review* is week-endwerk’, geeft een Belgische neuroloog toe. ‘Op zondag anderhalf uur een paper doornemen: dat is de realiteit. Het is ook vrijwillig: je doet het omdat je vindt dat het bij je taak van wetenschapper hoort.’ Er is dus – door de werk- en publicatiedruk – meestal geen tijd en geen stimulans om het goed en grondig te doen. *Peer review* is feilbaar, dat bleek toen een redacteur van *British Medical Journal* een artikel met acht duidelijke

lijk evengoed voor de herhaalstudies’, zegt Nuijten. ‘Er zijn aanwijzingen dat ook voor herhaalstudies geldt dat ze sneller gepubliceerd worden als ze significante resultaten bevatten.’ Ook onderzoekers zelf achten een oorspronkelijk onderzoek vaak gezaghebber dan hun eigen herhaling. Ze redeneren dan gauw dat het wel aan de uitvoering van het herhaalexperiment zal liggen als niet hetzelfde resultaat uit de bus komt, en gaan er niet op door. *Publication bias* verstoort ten slotte de meta-analyses of overzichtsstudies. Die worden als gezaghebbend beschouwd, omdat ze vele studies samennemen en ze individueel ‘wegen’, waarna een eindconclusie volgt. Maar de meta-analyse houdt vaak alleen rekening met de gepubliceerde studies, die positief zijn. De studies met nulresultaten, waarin vaktijdschriften niet geïnteresseerd zijn, zijn niet meegerekend. En dus overschat de meta-analyse dan het effect, en geeft ze een te rooskleurig beeld.

fouten in het studieontwerp, de analyse en de interpretatie van de resultaten naar meer dan 200 ervaren nalezers van het medische vakblad stuurde. Geen enkele pikte alle acht fouten eruit, en gemiddeld zagen ze er minder dan twee. In een ander experiment van *BMJ*, waarbij de nalezers vooraf op mogelijke problemen werden gewezen, brachten ze het er niet beter af. Ook groepsdruk kan het kritische oog blind maken. *Peer reviewers* zijn vakgenoten, ze lopen elkaar tegen het lijf op congressen, willen vaak nog samenwerken en zijn soms afhankelijk van elkaars oordeel voor het binnenhalen van fondsen. Om dezelfde redenen verlenen redacteurs van vaktijdschriften soms makkelijk doorgang aan *papers* van bevriende vakgenoten, en stellen ze bijvoorbeeld jonge, minder kritische *peer reviewers* aan. Nobelprijswinnaar Geneeskunde Sydney Brenner liet zich dit jaar in een interview ontvallen: ‘Natuurlijk zeggen alle academici dat we *peer review* moeten hebben, maar ik geloof er niet in. Het is een corrupt systeem geworden, het leidt tot vervlakking, en het hindert wetenschap.’

Zonder eind

De *eye-openers* van Bohannon en co doen wetenschappers steeds meer nadenken over een nooit eindigend reviewproces na de publicatie. Dat zoiets werkt, bewijst een discussie in februari over twee baanbrekende artikelen in *Nature*. Daarin beschreven Japanse onderzoekers hoe ze van volwassen muizencellen pluripotente stamcellen maakten door ze in een zuurbad te stoppen. Eenvoudig, maar misschien té eenvoudig. Na de publicatie verscheen op de website PubPeer harde kritiek van kenners, die vermoeden dat foto’s van experimenten gemanipuleerd werden. Het leidde er uiteindelijk toe dat een van de auteurs de *Nature*-paper wil terugtrekken, omdat hij er niet meer in gelooft. Een eerste herhaalpoging eind mei in *F1000Research* – een tijdschrift dat alle reviews openbaar maakt – bleek ook negatief. In 2011 moest een NASA-onderzoeker die in *Science* claimde dat arseen een basisbouwsteen voor leven is, zich verantwoorden na commentaar van vakgenoten op blogs. ‘Ook die commentaren op papers zouden overzichtelijk bij de online versie van de paper moeten staan’, vindt Ivan Oransky, oprichter van de website RetractionWatch, een waakhond-blog voor onderzoeksfraude. Dat *peer review* vandaag niet ‘open’ is, begrijpt Jelte Wicherts (Tilburg University) niet. ‘Het is vreemd dat een van de belangrijkste onderdelen van het wetenschappelijke proces achter gesloten deuren plaatsvindt’, schrijft hij in een paper uit 2012. Nu blijven nalezers vaak anoniem, en weet een lezer van een paper niet wat in de commentaren stond. Openheid over de commentaren en over wie de *reviewers* waren, zou de kwaliteit van het nalezen

kunnen verhogen. Zwakke *reviews* worden gespot, kwalitatieve commentaren beloond. En als *reviewers* officieel scores krijgen voor goed en kritisch geleverd werk, zoals Wicherts voorstelt, staat dat mooi op hun cv, of kunnen ze opklimmen tot redacteur van een vakblad. Nu ontbreken die stimulansen.

Op het platform Open Science Framework zijn al 300 onderzoekers geregistreerd die ‘open onderzoek’ doen

Brian Nosek wil zelfs nog verder gaan: hij wil ook het hele onderzoeksproces – van idee over mislukte experimenten tot resultaten – transparant maken. Hij heeft er een praktisch werkplatform voor opgezet: Open Science Framework, waar intussen een 300-tal onderzoekers op zijn geregistreerd. Het idee: onderzoekers worden betrouwbaarder omdat iedereen kan meekijken, én dergelijke platforms houden ook de negatieve resultaten boven de tafel. Minder druk op onderzoekers om alleen spectaculaire resultaten te produceren en openheid over de feilbaarheid van wetenschap, daarmee hoopt een groeiende groep vooruitstrevende wetenschappers de replicatiecrisis aan te pakken. Einstein wist het al: *All our science, measured against reality, is primitive and childlike – and yet it is the most precious thing we have.* ■

