

Hun ogen bedrogen hen zo sterk dat ze echt dachten, en vooral *voelden* dat hun vinger korter was geworden **W12**

NRC-SPECIAL

WETENSCHAP

Zaterdag 9 april & Zondag 10 april 2016



Puinruimen in de psychologie

Psychologen vliegen elkaar in de haren, nu veel experimenten niet herhaalbaar blijken. Ook in andere wetenschappen heersen twijfel en verontwaardiging. Een **special** over de onderzoekscrisis. **W4-11**

Psychologie

Er woedt een harde strijd in de psychologie over het herhalen van onderzoek. Een generatiekloof tekent zich af.

Door **Ellen de Bruin**

Er smeult iets in de wetenschap dat met regelmaat heftig opvlamt - en als dat gebeurt, dan gebeurt het het duidelijkst in de psychologie.

Zoals begin maart. In een artikel in *Science* schreven vier Amerikaanse onderzoekers, dat als je een experiment uit de psychologie opnieuw doet de kans groot is dat er hetzelfde uitkomt. Impliciete conclusie: het is dus een betrouwbare tak van wetenschap. Zo'n statement was nodig, vonden ze, na een schokkend artikel van een groep van honderden wetenschappers in augustus 2015, ook in *Science*. Die groep, Open Science Collaboration (OSC), had honderd gepubliceerde psychologische experimenten overgedaan en uit maar zo'n twee op de vijf was hetzelfde gekomen als uit het eerdere experiment. Het had wereldwijd het nieuws gehaald, ook *NRC*.

Maar dat replicatieproject van het OSC was dus heel slecht, schreven de vier Amerikanen nu: psychologische experimenten zijn juist goed repliceerbaar. Nietes, reageerden onmiddellijk 44 onderzoekers namens het OSC in dezelfde editie van *Science*. Waarna de ruzie zich verplaatste naar internet en daar, kort samengevat, als volgt verliep: „Welles. Echt, welles!” „Nietes!”

Straks meer hierover. Eerst: deze discussie onder psychologen (via blogs, tweets, Facebook en wetenschappelijke artikelen) is fundamenteel. Sommige psychologen maken zich weliswaar zorgen dat de status van hun vakgebied te grabbel wordt gegooid door dit soort ruzies publiekelijk uit te vechten, waarbij zichtbaar is dat onderzoekers fouten maken. Maar anderen vinden dat dit gesteggel over wat repliceerbaar is hoort bij goede wetenschap. Of ze zijn openlijk blij dat hun vakgebied het replicatieprobleem eindelijk aanpakt. Want nooit eerder in de ruim honderdjarige geschiedenis van de psychologie is er zo veel en systematisch replicatieonderzoek gedaan als nu. En ook in andere wetenschappen komt dit niet in die mate voor.

Woordspelletjes

Er wordt nu getwist over de vraag wat het betekent als een bepaald verschijnsel dat iemand eerst wél in onderzoek aantoonde, vervolgens niet meer te vinden is. Bij één replicatiepoging is er nog niet zoveel aan de hand. Neem een beroemd *priming*-experiment van John Bargh van Yale (*Journal of Personality and Social Psychology*, 1996), waarin mensen langzamer gingen lopen nadat ze een woordspelletje hadden gedaan waarin veel woorden voorkwamen die met het stereotype van ouderen te maken hebben ('eenzaam', 'grijs', 'bingo'). Als een ander onderzoeksteam dat effect niet vindt (zoals in een artikel in *PLOS One*, 2012), kan dat betekenen dat het niet bestaat. Maar het kan ook zijn dat het alleen onder specifieke omstandigheden optreedt, of dat de nieuwe onderzoekers het experiment net iets anders hebben uitgevoerd. En zowel het oorspronkelijke als het nieuwe resultaat kan ook nog *allebei* toeval zijn. Maar dit type onbewuste gedragspriming werd vervolgens vaker niet gerepliceerd.

Psychologische oorlogvoering



FOTO: ARJEN BORN

Het staat nog altijd met Barghs naam erbij in de lesboeken, maar steeds meer psychologen twijfelen eraan.

Net als aan *ego-depletion*, het gevestigde idee dat wilskracht uitgeput kan raken zoals een spier. Dat werd niet gerepliceerd in een groot project van 23 universiteiten (dit jaar) én ook niet teruggevonden in een meta-analyse (2015). Zo'n analyse neemt niet alleen het gepubliceerde onderzoek mee waar het effect uitkwam, maar houdt ook rekening met ongepubliceerd onderzoek waar het *niet* uitkwam. Dan wordt de twijfel wel erg groot.

In een emotionele blogpost schreef Michael Inzlicht (Universiteit van Toronto) eind februari dat hij bijna tien jaar onderzoek aan ego-depletion had gedaan, hij was ervoor bekroond, hield er lezingen over. „Heb ik al die jaren rookwolkjes nagejaagd?” Hij vraagt zich af welke gevestigde ideeën er nog meer onderuit zullen gaan. *Stereotype threat* misschien? Meisjes presteren slechter op wiskundetoetsen (waar ze volgens het stereotype slecht in zijn), als ze net subtiel herinnerd zijn aan hun geslacht - maar een meta-analyse uit 2015 suggereert dat ook dat effect misschien niet zo robuust is als werd gedacht. „Ik voel me alsof de grond onder me vandaan schuift en ik niet langer weet wat echt is en wat niet”, blogt Inzlicht, die een boek over stereotype threat redigeerde (het kwam in 2011 uit). Hij verwacht dat stereotype threat in een speciaal replicatieproject verder zal worden onderzocht. En ook dat zal dan wel weer ruzie opleveren.

Want zijn die replicatieprojecten zélf wel goede wetenschap? Daar wordt ook aan getwijfeld. Terug naar het meningsverschil in *Science* tussen het OSC en de vier Amerikanen, met sociaal psycholoog Daniel Gilbert (Harvard) als eerste auteur en het OSC. Volgens Gilbert en collega's heeft het OSC allerlei fouten gemaakt in de methode en analyses van hun omvangrijke replicatieproject. Zo had dat niet genoeg statistische 'power' om goede conclusies te kunnen trekken, vonden Gilbert c.s.: elk experiment werd maar één keer overgedaan. Verder ging het louter om relatief makkelijk uit te voeren experimenten uit drie psychologietijdschriften: geen willekeurige steekproef. De OSC-onderzoeksteams mochten bovendien kiezen wat ze over wilden doen, ze werden niet random aan onderzoek toegewezen. Dat vergroot de mogelijkheid dat sceptische onderzoekers de resultaten negatief beïnvloedden.

En het OSC heeft de experimenten volgens Gilbert c.s. niet exact genoeg herhaald - altijd een kwestie, bij replicaties. Hierover wordt de discussie extra onaangenaam als Gilbert c.s. een experiment zo vertekend beschrijft dat het OSC geblunderd lijkt te hebben, wat de groep dan weer rechtzet op een blog. Zo is de hele discussie een uitwisseling van 'ja maar jij deed dit, ja maar jij deed dat', waarbij toponderzoekers - het zijn net mensen - fouten maken, statistische en methodologische meningsverschillen hebben en informatie bevooroordeeld beschrijven.

Intuïtieve flair

Immiddels hooft de sociaal-psychologische gemeenschap alweer over iets nieuws: ego-depletion-pionier Roy Baumeister, die in een binnenkort te verschijnen replicatiethemanummer van *Journal of Experimental Social Psychology* onder meer klaagt dat er steeds grotere proefpersoon aantallen worden geëist. „Een significant resultaat verkrijgen met n=10 [tien proefpersonen, red.] vereiste vaak een intuïtieve flair voor hoe je de meest gunstige situatie moest inzetten om een uiterst krachtige procedure te produceren”, schrijft hij over zijn jonge jaren.

Vakgenoten reageren lacherig: flair inzetten om gewenste resultaten te krijgen? Dat klinkt wetenschappelijk dubieus.

De vraag wat er precies repliceerbaar is in de psychologie en wat de uitkomsten van replicaties betekenen, is al met al nog steeds niet definitief beantwoord. En het is al een oud probleem. In 1976 schreef de toenmalige hoofdredacteur van *Journal of Personality and Social Psychology*, nog altijd een van de belangrijkste tijdschriften in het vakgebied, bij zijn aantreden: „Er is misschien een crisis in de persoonlijkheids- en sociale psychologie die verband houdt met de moeilijkheid die onderzoekers vaak ervaren in hun pogingen om gepubliceerd werk te repliceren.” Hij eiste dat auteurs alle procedurele informatie die nodig was om onderzoek te repliceren tegelijk met het manuscript instuurden. Ook moesten auteurs meer statistische informatie meesturen dan afgedrukt kon worden. De maatregelen werden later teruggedraaid; zorgelijke en minder zorgelijke fases in de wetenschap wisselen elkaar af.

De zorgelijke periode waar de psychologie nu middenin zit, heeft verschillende aanleidingen. Voor het grote publiek was de grootschalige datafraude van Diederik Stapel, die in 2011 aan het licht kwam, de duidelijkste. Psychologen noemen ook vaak een artikel uit datzelfde jaar, waarin Daryl Bem (Cornell) 'aantoonde' dat mensen in de toekomst kunnen kijken: zijn proefpersonen leken boven kansniveau te kunnen voorspellen waar op het computerscherm een erotische afbeelding zou verschijnen (*Journal of Personality and Social Psychology*, maart 2011). Als je dat kunt aantonen, is er iets mis met de gebruikte statistiek, schreven vier Amsterdamse psychologen die Bems artikel in hetzelfde tijdschrift afbrandden. Ze lieten zien dat Bems effecten verdwenen met conservatievere (Bayesiaanse) toetsen.

Revolutie 2.0

Vanaf die tijd raakten steeds meer psychologen ongerust: over te vrijmoedige statistische methoden, te weinig replicaties, fraude en andere dubieuze onderzoekspraktijken. Die problemen worden met enige regelmaat onderkend. Maar dit keer is het anders, schreef rechtspsycholoog Barbara Spellman (Universiteit van Virginia) in *Perspectives on Psychological Science* (2015), toen ze er afrad als hoofdredacteur. Zij noemt de huidige „introspectieve fase” in de psychologie geen crisis, maar een wetenschapspolitieke revolutie: 'Revolutie 2.0', vanwege de cruciale rol voor nieuwe technologie.

Nog maar enkele decennia geleden moesten psychologen voor elke statistische test die ze wilden doen een opdracht geven aan de enige computer op de vakgroep die zo iets aankon. Artikelen moesten ze bij de auteurs opvragen als de universiteit niet geabonneerd was op het tijdschrift. Die stuurden dan een papieren kopie per post. Manuscripten moesten ook op papier, in viervoud, naar tijdschriften worden gezonden; vragenlijsten werden op papier afgenomen. Nu draait iedereen elke gewenste statistische test in een paar seconden op de eigen computer, is elk artikel en elke collega waar ook ter wereld een paar muisklikken verwijderd en kunnen onderzoekers duizenden proefpersonen op één dag een vragenlijst laten invullen, waarna de data al 'in de computer zitten'.

Onverwacht gevolg van die technologische vernieuwingen, schrijft Spellman, is dat mensen nu veel sneller horen dat het een ander lab óók niet is gelukt om een onderzoek te repliceren. Vroeger hoorden onderzoekers zo iets toevallig, 's avonds laat op congressen. Jonge promovendi



Heb ik al die jaren rookwolkjes nagejaagd?

Psycholoog Michael Inzlicht

ploeterden soms jaren om een effect te repliceren waar ze op wilden voortbouwen. Als dat niet lukte, hadden ze niets. Nu zoeken ze direct contact met andere groepen. De bewijslast dat zo'n effect echt bestaat is daardoor eerder bij de oorspronkelijke onderzoekers komen te liggen dan bij de onervaren promovendus.

Dat hangt samen met een andere verandering die Spellman beschrijft: doordat het aantal jonge onderzoekers de laatste decennia enorm is gegroeid, is de kracht van het *old boys' network* sterk afgenomen. In feite is er een *young researchers' network* van *digital natives* naast gezet. De huidige meningsverschillen in de psychologie worden ruwweg uitgevochten door een 'jong' en een 'oud' kamp, waarbij vooral de jongere generatie pleit voor nieuwe statistiek, meer replicaties, vooraf registreren van hypothesen en methoden van onderzoek, open access (alle artikelen gratis online voor iedereen beschikbaar) en open wetenschap in het algemeen (alle data bij alle artikelen online beschikbaar). Sommige psychologen (vooral oudere) vinden dat de vernieuwers doorslaan. Zoals dus Roy Baumeister, die vreest dat door de nadruk op replicaties en zorgvuldigheid alle creatieve innovatie uit het vakgebied verdwijnt.

Incompetente onderzoekers

De ruziënde partijen blijven niet altijd beleefd. De eerste reactie van Bargh op de onderzoekers die zijn langzaam-lopen-experiment niet konden repliceren, was dat het kon liggen aan „incompetente of slecht geïnformeerde onderzoekers”, of aan de inferieure *peer review* die open access-tijdschrift PLOS One volgens hem had. En Gilbert heeft replicatie-onderzoekers weleens „schaamteloze kleine pestkoppen” genoemd. Dat was toen die onderzoekers vergeefs een experiment van Simone Schnall (Cambridge) probeerden te repliceren en daar op het aggressive af enthousiast over blogden („*an epic fail as my 10 year old would say*”).

Dat bekechten over replicaties trekt de aandacht, maar er zijn ook andere belangrijke problemen. Bijvoorbeeld: doordat onderzoekers afgerekend worden op publicaties, is het gunstig voor ze om hun onderzoek zo versnipperd mogelijk te publiceren - maar dat is slecht voor de kennisopbouw. En hoewel internetruimte vrijwel onbeperkt is, vragen tijdschriften toch om de methodesecties kort te houden - dat maakt repliceren moeilijker.

Volgens Spellman verandert dat systeem niet snel doordat vooral de oudere generatie de tijdschriften en subsidiepotjes nog beheert. Toch verwacht hij dat de revolutie slaagt, omdat het tijdschriftologisch meezit. Het publicatiesysteem kan op de schop: alle studies kunnen op internet vooraf geregistreerd worden, alle methoden en alle data kunnen op internet, net als studies die onpubliceerbaar zijn omdat er niks uitkomt, maar waarvan het wel interessant is dat er niks uitkomt. Open en vrije uitwisseling van kennis, het oude ideaal, wordt dan de nieuwe norm waar vanuit de wetenschap verder kan.

Ironisch genoeg is Spellman óók optimistisch over Revolutie 2.0 omdat de problemen zich niet tot de psychologie beperken. De huidige crisis, of revolutie, is van de hele wetenschap. Allerlei vakgebieden kampen met replicatieproblemen en hebben te maken met een nieuwe generatie *digital natives*. En qua fraude: Diederik Stapel hoort met 58 ingetrokken artikelen niet bij de top-3 op weblog Retraction Watch. Die bestaat uit een Japanse en een Duitse arts en een Taiwanese fysicus. Maar psychologen voeren de discussies over hervormingen in de wetenschap veel publieker dan andere vakgebieden. En hoe dat dan komt? Ook dat weten we niet.