Good science, bad science: Questioning research practices in psychological research

Bakker, M.

Nederlandse samenvatting

Goede en Slechte Wetenschap.

Betwistbare Onderzoekspraktijken in de Psychologische Wetenschap

De afgelopen jaren is er veel discussie geweest over de betrouwbaarheid van psychologisch onderzoek. Deze discussie is opgelaaid naar aanleiding van een aantal gebeurtenissen, zoals de ontdekking dat de sociaal psycholoog Diederik Stapel data had verzonnen. Iedereen is het erover eens dat dit soort fraude niet mag. Daarom is het interessanter om te kijken naar gedrag van onderzoekers in het grijze gebied tussen fraude en goede wetenschap. Een interessantere gebeurtenis is daarom de publicatie van Daryl Bem in een toonaangevend sociaal psychologisch tijdschrift, waarin hij met meerdere studies aantoont dat mensen in de toekomst kunnen kijken. Dit roept de vraag op hoe het kan dat zo'n ongeloofwaardig resultaat gevonden en gepubliceerd kan worden.

Om dieper op deze vraag in te gaan zal ik eerst nulhypothesetoetsing kort uitleggen. Dit is een toetsingsmethode die veel gebruikt wordt in psychologisch onderzoek. Bij deze methode wordt begonnen met een nulhypothese van geen verschil of geen effect. Stel dat een onderzoeker bijvoorbeeld wil onderzoeken of mensen die een mok met warme koffie vasthouden andere mensen vervolgens ook als 'warmer' beoordelen. Om dit te onderzoeken kun je een groep proefpersonen willekeurig in twee groepen verdelen. In de ene groep houden alle deelnemers kort een mok met warme koffie vast en in de andere groep niet. Vervolgens moeten alle proefpersonen dezelfde fictieve persoon beoordelen op onder andere 'warmte'. De nulhypothese is in dit geval dat er *geen* effect is van het wel of niet vasthouden van de mok op de beoordeling van de 'warmte' van een fictief persoon. De gemiddelde score op 'warmte' in de ene groep wordt vervolgens vergeleken met de gemiddelde score in de andere groep en dit verschil wordt samengevat in een toetsingsgrootheid. Daarna wordt gekeken wat de kans is op de gevonden toetsingsgrootheid (of nog extremer) gegeven dat de nulhypothese waar is. Hoe verder de gemiddelden van de twee groepen uit elkaar liggen, hoe kleiner deze kans is. Deze kans wordt de *p*-waarde genoemd; wanneer deze lager is dan de afgesproken waarde (meestal .05) mag de nulhypothese verworpen worden en wordt de alternatieve hypothese

geaccepteerd. In dat geval nemen we aan dat er een effect is (het vasthouden van een mok met warme koffie zorgt ervoor dat men een fictief persoon als 'warmer' beoordeelt). Omdat het verwerpen van de nulhypothese de kans op een publicatie flink vergroot, is het voor onderzoekers zeer belangrijk dat de *p*-waarde lager dan .05 is.

Nu kunnen tijdens nulhypothesetoetsing twee soorten statistische fouten worden gemaakt. De eerste is de Type I fout (fout-negatief). Hierbij is de nulhypothese waar (er is geen effect) en toch vinden we dat $p < .05$ is en accepteren we onterecht de alternatieve hypothese. Omdat we hebben afgesproken dat we bij $p < .05$ de nulhypothese mogen verwerpen, is de kans op een Type I fout 5% als de nulhypothese waar is. De tweede fout is de Type II fout (fout-positief) en daarbij is het precies andersom. De alternatieve hypothese is waar (er is een effect), maar we kunnen op grond van de data de nulhypothese niet verwerpen. Vaak willen onderzoekers dat de kans op deze fout maximaal 20% is. Anders gezegd, ze willen een onderscheidingsvermogen (*power*) van .8 hebben om als het effect bestaat dit ook aan te kunnen tonen met het experiment.

De meeste mensen zullen ervan overtuigd zijn dat we niet in de toekomst kunnen kijken en dat dus de nulhypothese in het onderzoek van Daryl Bem waar is. Toch heeft hij deze nulhypothese kunnen verwerpen. Mogelijk heeft hij heel vaak 'geluk' gehad en meerdere keren een Type I fout gevonden. Het zou ook kunnen dat hij hetzelfde onderzoek vaak heeft herhaald, alleen de 'gelukte' studies gepubliceerd heeft en de rest van de studies in de kast laat liggen. Een andere mogelijkheid is dat hij gebruik heeft gemaakt van betwistbare onderzoeksmethoden (*questionable research practices*). Dit zijn methoden die in het grijze gebied vallen tussen fraude en goed onderzoek en mogelijk de kans op een Type I fout vergroten. Het zijn ook vaak methoden die in de ene context wel te verdedigen zijn, maar in de andere niet.

John, Loewenstein en Prelec (2012) hebben onderzocht hoe vaak een aantal van deze betwistbare onderzoeksmethoden voorkomt. Een van de de betwistbare onderzoeksmethoden uit dat onderzoek is het incorrect afronden van *p*-waardes. Oftewel, *p*-waardes net boven de .05 rapporteren alsof ze net onder deze belangrijke grens liggen. Dit werd toegegeven door 22% van de onderzoekers die de vragenlijst van John et al. hebben ingevuld. Deze afrondingsfouten kun je herkennen in artikelen door de *p*-waarde te bereken op basis van de gerapporteerde toetsingsgrootheid (bijvoorbeeld de *t* waarde) en het gerapporteerde aantal vrijheidsgraden, en deze te vergelijken met de gerapporteerde *p*-waarde. Wij hebben dit gedaan voor een groot aantal artikelen en we vonden dat ongeveer de helft van deze artikelen minimaal één zo'n rapportagefout bevat. Dat kunnen kleine afrondingsfouten zijn, maar in 15% van de artikelen zat een fout die de statistische conclusie veranderde. Dit staat in hoofdstuk 2 van dit proefschrift. In hoofdstuk 3 hebben we vervolgens onderzocht of onderzoekers die hun data delen, zoals van

een goede onderzoeker verwacht mag worden, ook minder fouten maken. Dit bleek inderdaad het geval te zijn. Daarnaast liggen de significante *p*-waardes in de artikelen waarvan de data niet gedeeld waren dichter bij de grenswaarde van .05, wat kan worden gezien als minder sterk bewijs tegen de nulhypothese in de artikelen waarvan de data niet gedeeld werden.

Een andere betwistbare onderzoeksmethode is het verwijderen van uitbijters (extreme datapunten of *outliers*) uit de data, omdat zowel het verwijderen als het laten zitten van uitbijters een grote impact kan hebben op de uitkomsten van een analyse. 38% van de onderzoekers die de vragenlijst van John et al. hadden ingevuld gaven toe dat ze hadden besloten om data te verwijderen nadat ze hadden gekeken wat de impact van het verwijderen op de uitkomst was. Daarom onderzoeken we in hoofdstuk 4 of in artikelen waarin uitbijters zijn verwijderd ook meer fouten worden gemaakt. Dit kunnen we niet aantonen. Een mogelijk probleem hierbij is dat onderzoekers niet altijd precies bijhouden of er data is verwijderd. In hoofdstuk 5 gaan we verder in op het verwijderen van uitbijters uit de data en de invloed die dit heeft op de Type I fout van de *t*-test. Uit simulatiestudies blijkt dat een veel toegepaste methode om uitbijters te verwijderen ervoor zorgt dat de Type I fout een stuk hoger is dan de afgesproken vijf procent. Zeker als de data daarnaast ook niet normaal verdeeld zijn, of als de methode subjectief wordt toegepast. Non-parametrische en robuuste methoden zijn minder gevoelig voor de aanwezigheid van uitbijters en zijn daardoor meer geschikt wanneer de data mogelijk uitbijters bevatten.

In hoofdstuk 6 onderzoeken we waarom onderzoekers veel kleine studies doen met weinig onderscheidingsvermogen en/of betwistbare onderzoeksmethoden gebruiken. Daarvoor hebben we een simulatiestudie uitgevoerd waarbij we de wetenschap benaderen als een spel waarin het doel is zoveel mogelijke publicaties te krijgen (wat belangrijk is voor een carrière als wetenschapper). Om dit spel te winnen zal de onderzoeker zoveel mogelijk significante ($p < .05$) resultaten moeten vinden. Wij laten zien dat de beste strategie is om veel kleine studies te doen en daarbij gebruik te maken van de betwistbare onderzoeksmethoden. Hiermee is de kans op een significant resultaat het grootst. Helaas heeft dit ook als gevolg dat veel Type I fouten gemaakt worden en de effectgrootte overschat wordt. Voor de wetenschap als geheel is deze strategie dus funest. In hoofdstuk 7 presenteren we ons onderzoek naar een andere reden waarom onderzoekers vaak veel kleine studies met maar heel weinig onderscheidingsvermogen doen. Om te kijken of intuïties van onderzoekers over het onderscheidingsvermogen van hun onderzoek wel kloppen, hebben we onderzoekers gevraagd om hun typische steekproefgrootte, effectgrootte, α (de grenswaarde voor *p*) en het gewenste onderscheidingsvermogen op te geven. Hieruit bleek dat onderzoekers meestal een onderscheidingsvermogen willen van .8, maar als we het onderscheidingsvermogen berekenen op basis van de steekproefgrootte, effectgrootte en α is het onderscheidingsvermogen maar ongeveer de helft.

De hoofdstukken uit dit proefschrift leveren een bijdrage aan de huidige discussie over de kwaliteit van wetenschappelijk onderzoek in de psychologie (en andere vakgebieden). Deze discussie is sinds 2011 opgelaaid, maar is ook eerder al gevoerd. Wat deze keer anders is, is dat de huidige discussies echt invloed lijken te hebben op de praktijk. Dat ligt waarschijnlijk ook aan de mogelijkheden die techniek en internet tegenwoordig bieden om data te bewaren en te delen. Veel nieuwe initiatieven worden gestart, zoals meer aandacht voor replicatie en transparantie; hopelijk lijdt dit tot structurele verbeteringen binnen de psychologische wetenschap.