Open Science Collaboration (2015) rapporteert de resultaten van replicatieonderzoeken van honderd eerder gepubliceerde onderzoeksresultaten uit vooral de cognitieve en sociale psychologie. Van deze resultaten waren 97 positief (significant, p<.05). Er werden dus effecten gerapporteerd, maar slechts 36% hiervan bleek bij replicatie ook positief te zijn. De Volkskrant *en* NRC *uitten onmiddellijk hun zorg over de waarde van psychologisch onderzoek.*
Klaas Sijtsma over replicatiestudies en het raadplegen van methodologen.

RAADPLEEG EERST EEN METHODOLOOG!

HEEFT DE PSYCHOLOGIE REPLICATIE-STUDIES NODIG?

De Tilburgse methodoloog Michèle Nuijten legde in *de Volkskrant* (5 september 2015) helder uit hoe we zulke percentages moeten interpreteren. Zij riep terecht op tot het gebruik van grotere steekproeven en openbaarheid van onderzoeksdata om nepeffecten – dit zijn effecten die wel worden gevonden maar geheel te wijten zijn aan de toevallige samenstelling van de steekproef – te voorkomen.

In dit artikel vraag ik aandacht voor een derde remedie: het bij onderzoek tijdig inschakelen van een methodoloog. Ik bespreek eerst dat methodenleer en statistiek voor veel onderzoekers te moeilijk zijn om 'er even bij te doen', en dat het daarom verstandig is een methodoloog in te schakelen. Vervolgens kom ik terug op het onderzoek van Open Science Collaboration (2015) en zal ik beargumenteren dat, hoe nuttig en goedbedoeld ook, dit onderzoek hooguit een eerste aanzet is.

'METHODENLEER EN STATISTIEK DOE JE ER NIET EVEN BIJ'

Psychologen en andere onderzoekers hebben methodenleer en statistiek nodig om hun onderzoek te kunnen doen, maar ze onderkennen vaak te weinig dat zij dit moeilijke vak onvoldoende beheersen. Deze onderschatting leidt tot vele en soms ernstige problemen die bekend staan als *questionable research practices* (QRP's). Een voorbeeld van een QRP is het stelselmatig weglaten van waarnemingen uit de data die het vinden van een gewenst resultaat, veelal een significant effect, in de weg staan. Dit komt erop neer dat men net zolang proefpersonen uit de steekproef weglaat en anderen weer opneemt, totdat het gewenste resultaat wordt behaald. Niet de gehele steekproef bepaalt het resultaat, maar de selectie die de onderzoeker hieruit maakt waarin de toets significant is. Je kunt dit vergelijken met een boogschutter die zijn pijl afschiet, daar vervolgens de roos omheen schildert en vervolgens roept dat hij in de roos heeft

geschoten. Uiteraard staan de spelregels van de methodenleer dit niet toe, zoals ook de beroemde psycholoog Adriaan de Groot ruim vijftig jaar geleden in zijn klassieker *Methodologie* overtuigend uiteenzette (De Groot, 1961; zie ook Busato, 2014).

Men kan denken dat methodologische spelregels zelden overtreden worden, maar als het doel het rapporteren van mooie resultaten is, dan worden de middelen wel eens aangepast. Dit is niet eens altijd opzet, en dan weet men dus eenvoudig niet dat deze manier van werken weinig bijdraagt aan waarheidsvinding. Het onderzoek naar de vele soorten QRP's is een apart vakgebied geworden (Bakker, Van Dijk & Wicherts, 2012; Simmons, Nelson & Simonsohn, 2011), dat laat zien dat QRP's vaak voorkomen (John, Loewenstein & Prelec, 2012).

De lezer krijgt alleen de succesverhalen te zien maar niet de vergeefse pogingen een effect aan te tonen

NEPEFFECTEN KOMEN VEEL VOOR

Door allerlei methodologische en statistische fouten wordt er veel onderzoek gedaan naar nepeffecten (Ioannides, 2005). Een denkbeeldig voorbeeld is onderzoek naar een ernstige ziekte, zeg A (ik houd het bij A om te voorkomen dat de lezer bij een concrete ziekte denkt dat het onderzoek ernaar niet zou deugen). Onderzoek naar de oorzaak van A en de remedie ertegen is zeer aantrekkelijk. Wat gemakkelijk gebeurt, is dat men op basis van een kleine steekproef concludeert dat een toevallig gevonden effect echt is en dit resultaat zo snel mogelijk publiceert. Andere onderzoekers lezen de publicatie, zijn enthousiast en gaan zich ook met het onderzoek naar ziekte A bezighouden. Hierdoor ontstaat concurrentie, die haast en slordigheid in de hand werkt, zodat het nog lang kan duren voordat het oorspronkelijk gevonden nepeffect als zodanig wordt ontmaskerd.

Men kan nu gemakkelijk tegenwerpen dat onderzoekers niet zo onnozel zijn dat ze niet door hebben dat ze een nepeffect hebben gevonden. De literatuur over QRP's laat evenwel zien dat een heel conglomeraat van elkaar versterkende factoren het voor onderzoekers moeilijk maakt zich aan de concurrentieslag te onttrekken, waardoor het nepeffect niet snel als zodanig wordt herkend. Om te beginnen zijn onderzoekers ingesteld op het vinden van effecten en zouden zij graag een grote en belangrijke vondst publiceren. Het verlangen effecten te vinden werkt in de hand dat men confirmerend te werk gaat en het onderzoek niet zo inricht dat verwachtingen ook ontkracht kunnen worden als ze onjuist blijken te zijn. Confirmatie heeft dus de overhand en het Popperiaanse idee (Popper, 1935; 2002) dat je als onderzoeker moet proberen je eigen theorie onderuit te halen – falsificatie –, wordt vaak niet nagestreefd.

Ook wetenschappelijke tijdschriften zijn op confirmatie gericht en publiceren dus liever positieve dan negatieve resultaten, in de kennelijke veronderstelling dat niemand geïnteresseerd is in onderzoekrapportages waarin een verwacht effect niet werd gevonden. Een artikel waarin een effect van een therapie voor ziekte A wordt aangetoond, wordt dus eerder gepubliceerd dan een artikel waarin dit effect niet kon worden aangetoond. Door negatieve resultaten niet te publiceren, maken tijdschriften zich schuldig aan 'publication bias'. De lezer krijgt alleen de succesverhalen te zien maar niet de vergeefse pogingen een effect aan te tonen. Onderzoekers weten dat tijdschriften zo handelen en leggen negatieve resultaten terzijde als niet publiceerbaar: het 'filedrawer effect'. Dit is een vorm van zelfcensuur, die bekend staat als 'selection bias'.

HAAST Deze selectiemechanismen werken in de hand dat een nepeffect de status krijgt van een echt effect. Niemand is immers op de hoogte van alle negatieve resultaten die de overhand hebben, maar die nooit zijn gepubliceerd. Maar er is meer. Ik noemde al haast als gevolg van de drang om positieve resultaten te behalen. Daardoor nemen onderzoekers een kleine steekproef voor lief omdat het verzamelen van meer gegevens tijd kost en men de concurrentie voor wil blijven. Kleine steekproeven laten geen nauwkeurige

schattingen van effectgrootten toe en door toeval kan een steekproefeffect groot zijn terwijl het in werkelijkheid om een nepeffect gaat.

Onderzoekers worden verder tot haast aangespoord door individuele motieven zoals het vooruitzicht op een mooi artikel, een grote subsidie, een prestigieuze prijs of een bevordering, en voor jonge onderzoekers hangt een vaste aanstelling vooral af van publicaties in goede tijdschriften. De kans op onjuiste conclusies uit onderzoek wordt nog eens dramatisch vergroot door de vele methodologische en statistische dwaalsporen die vaak onbewust bewandeld worden (John et al., 2012; Simmons et al., 2011). Verder wordt die kans vergroot door de gewoonte om bij een negatief resultaat, waar je een positief resultaat verwachtte, explorerend op zoek te gaan naar mogelijk andere, niet verwachte maar wel interessante resultaten in de data. Zo kan men toch iets publiceren, zij het iets anders dan gepland. Onderzoekers lopen dan een verhoogd risico om op kans te kapitaliseren, ofwel onbedoeld te profiteren van steekproeffouten, door te lang door te zoeken en nepeffecten als echte effecten te rapporteren.

OPEN SCIENCE COLLABORATION Replicatieonderzoek is een precieze nabootsing van het oorspronkelijke onderzoek met een nieuwe steekproef van proefpersonen, waardoor de resultaten van verschillende replicatieonderzoeken alleen verschillen door toevallige verschillen tussen de steekproeven. Zo krijg je zicht op de invloed van toeval op de resultaten en zou je nepeffecten in het oorspronkelijke onderzoek als zodanig kunnen ontmaskeren. Immers, als het effect niet echt bestaat, maar tot stand kwam door de toevallige samenstelling van de steekproef, dan verwacht je in een nieuwe steekproef dat precies die toevalligheden die het nepeffect veroorzaakten niet nogmaals voorkomen. Er gebeurt echter maar weinig replicatieonderzoek, vandaar dat het onderzoek van Open Science Collaboration (2015) vanwege de voorbeeldfunctie zo belangrijk is.

Wat in de onderzoekspraktijk wel gebeurt, is dat een nepeffect niet als zodanig herkend wordt. Het is echter wel aantrekkelijk, omdat het bijvoorbeeld een remedie tegen ziekte A belooft. Dat trekt al snel andere onderzoekers aan, van wie velen via slechte publicatiegewoonten en vele methodologische en statistische ongerijmdheden weer dezelfde fouten maken, waardoor het lijkt of het positieve effect gerepliceerd is. Maar er zijn natuurlijk ook onderzoekers die, als ze een veel kleiner effect of zelfs niets vinden, zich afvragen of het positieve effect geen toeval was. Zo ontstaan er na verloop van tijd ook kritische artikelen waarin alternatieve hypothesen worden geopperd en onderzocht en dooft de onderzoekshype naar ziekte A weer uit.

Psychologen zijn met het beschreven fenomeen bekend onder de naam 'regressie naar het gemiddelde' (Hand, 2014, p. 169). Dit is een berucht methodologisch artefact, dat ontstaat als je uit een groot aantal waarnemingen een of enkele extreme waarnemingen selecteert, die mede extreem zijn door toeval, en daardoor bij replicatie waarschijnlijk minder extreem uitvallen. Het gaat om een selectie-effect, dat je op vele plaatsen, zoals tentamens, kunt zien, maar dat slecht wordt begrepen. Voor het begrip van het effect is nodig dat je weet dat de correlatie tussen een toetsscore en toevallige meetfouten positief is; dit betekent dat hoge toetsscores gemiddeld samengaan met positieve meetfouten, zeg maar geluk, en lage toetsscores gemiddeld met negatieve meetfouten ofwel pech.

Als je een groep studenten dus indeelt in een groep met lage scores, de gezakten, en een groep met hoge scores, de geslaagden, dan heb je ze daarmee niet alleen op kennis en vaardigheid geselecteerd, maar ook op respectievelijk pech en geluk. Als de groep gezakten zonder zich voor te bereiden aan een even moeilijke herhaling zou meedoen, heeft een deel van de gezakten deze keer wel geluk. Pech en geluk vallen nu tegen elkaar weg en de groep haalt een hogere gemiddelde toetsscore dan de eerste keer, waardoor sommigen nu wel slagen. Ook een nepeffect is het gevolg van selectie: de onderzoeker zag een significant resultaat op een terrein waar hij dat graag wilde zien, negeerde alle niet-significante resultaten die hij misschien ook al had gezien en ging verder met dit geselecteerde, extreme resultaat. Bij een replicatiepoging valt het dan tegen.

GEEN OPZET, MAAR HEBBEN ONDERZOEKERS DAN NIETS DOOR?

Uit de uiteenzetting over nepeffecten blijkt nog niet meteen dat methodenleer en statistiek moeilijk zijn. Misschien is alles wel kwade opzet en publiceert men bewust nepeffecten om er beter van te worden. Maar Tversky en Kahneman hebben in een reeks van studies (bijv. Tversky & Kahneman, 1974) overtuigend laten zien dat onderzoekers niet statistisch kunnen denken en dat veel methodologen en statistici er ook moeite mee hebben. Dat laatste mag dan voor sommigen een troost zijn, maar daar schieten we niet veel mee op. Tversky en Kahneman verklaren de moeilijkheid van statistiek uit de grote moeilijkheid van statistisch redeneren, dat in veel gevallen ook nog eens tegen-intuïtief is.

REPLICATIE MILGRAM-EXPERIMENT

CARTOON: AREND VAN DAM

Een overtuigend voorbeeld is regressie naar het gemiddelde, dat ik net besprak. Probleem is dat het begrip toeval er in de vorm van pech en geluk een rol in speelt, en mensen zich met het begrip toeval vaak geen raad weten. Zo houden deelnemers aan loterijen er allerlei strategieën op na, terwijl hen wel degelijk wordt verteld dat het resultaat van de loterij door toeval bepaald wordt (Hand, 2014). Je kunt dus rustig aselect een getal kiezen, maar dat lijkt dommer dan het volgen van een uitgekiende strategie. Dat niets doen (aselect kiezen) net zo effectief is als een strategie volgen (namelijk: niet effectief), is op zich al tegen-intuïtief en maakt het begrijpen van toeval alleen maar lastiger.

Statistiek gaat over het in kaart brengen van onzekerheid en dus over kansen, en ook het denken in termen van kansen blijkt voor de meeste mensen een onmogelijke opgave. Vaak wordt de kans op een bepaalde gebeurtenis verkeerd getaxeerd. Wie hiervan niet overtuigd is, mag zijn intuïtie aanwenden om het antwoord op de volgende vraag te geven: Stel, ik zet 23 aselect gekozen mensen bij elkaar, wat is de kans dat minstens twee van hen op dezelfde dag jarig zijn? Google (Birthday Problem) en Ross (1994, p. 42) geven het antwoord[1]. Het narekenen van dit antwoord is nog niet zo eenvoudig, en daarom zie ik er hier vanaf. Mensen denken overigens altijd dat de kans klein is, omdat 23 mensen maar weinig is in vergelijking met 365 dagen. Ik raad de lezer aan dit probleem aan bekenden voor te leggen en na te gaan of dit klopt. Maar de intuïtie is dus onjuist en het correcte, tegen-intuïtieve antwoord vereist een flinke inspanning.

Tot slot noem ik een ander tegen-intuïtief fenomeen dat zich voordoet bij het stapsgewijs aanvullen van steekproeven, ook wel 'bijdraaien' genoemd. Dit houdt in dat men

1 De kans is net iets groter dan .5.

begint met een kleine steekproef, waarop de nulhypothese wordt getoetst. Gegeven dat onderzoekers significante resultaten wensen, wil men graag de nulhypothese verwerpen en concluderen dat er een effect is. Lukt dat niet, dan wordt het onderzoek voortgezet en worden bij enkele nieuwe proefpersonen data verzameld. Deze data worden aan de steekproef toegevoegd en de nulhypothese wordt opnieuw getoetst. Is de toets niet-significant, dan wordt er weer 'bijgedraaid', enzovoort. Men stopt als de nulhypothese verworpen wordt, rapporteert het significante resultaat en stuurt het artikel naar een tijdschrift. Deze procedure wordt in experimenteel onderzoek veel gebruikt, maar wat velen niet weten is dat het met volledige zekerheid, dus kans 1, leidt tot het verwerpen van de nulhypothese, zelfs als deze waar is. De procedure kent dus maar een uitkomst – de nulhypothese wordt altijd verworpen – en kan dus geen informatie geven over het onderzochte fenomeen (Kadane, Schervish & Seidenfeld, 1996). Deze en andere fouten geven aanleiding tot talloze onjuiste conclusies in het gebruik van methodenleer en statistiek.

Veel methodologen en statistici hebben ook moeite met statistisch denken

De voorbeelden betreffen alle problemen met tegen-intuitieve uitkomsten, en vereisen heel wat inspanning om de juiste antwoorden te vinden. Je moet maar net doorhebben dat er iets niet klopt aan je intuïtie en vervolgens wat dat zou kunnen zijn, en de resultaten zijn voor niet-ingewijden volstrekt onverwacht. Huff (1954), Abelson (1995), Kahneman (2011) en Hand (2014) geven overtuigende onderbouwingen van de moeilijkheid van statistisch denken, waarbij Huff en Hand zich concentreren op de slimme leek en Abelson en Kahneman op de onderzoeker. Een les uit Kahneman (2011) is dat onderzoekers die geen ervaren statistici zijn, veelal vertrouwen op hun intuïtie - die dus meestal niet klopt, terwijl statistici ook hun intuïtie gebruiken maar met het verschil dat die gebaseerd is op veel ervaring die juist maakt dat ze erg voorzichtig zijn. Een statisticus die het birthday problem niet zou kennen en er voor het eerst mee geconfronteerd wordt, denkt al gauw dat 23 personen heel veel verschillende paren bevatten die allen op dezelfde dag jarig kunnen zijn. Dan zegt je intuïtie je dat het probleem misschien wel wat lastiger is dan het op het eerste gezicht lijkt. Maar de ervaring met statistiek is dus nodig om meteen het verband met paren in plaats van individuen te leggen.

Onderzoekers beseffen meestal niet dat zij essentiële statistische kennis en ervaring missen en gaan ongemerkt af op hun intuïtie. Maar statistiek is vooral tegen-intuïtief, en dus gaat men voortdurend de fout in. Dit is wat de literatuur over QRP's laat zien, maar wat moeilijk te verkopen is, juist omdat ook de boodschap – statistiek is tegen-intuïtief en dus moeilijk – in strijd is met de intuïtie. Docenten die statistiek onderwijzen, zouden vooral veel aandacht moeten besteden aan het overbrengen van de boodschap dat statistiek meestal tot andere conclusies leidt dan je zou verwachten en dat het daarom een moeilijk vak is. Dit is een impopulaire boodschap, die gemakkelijk verkeerd kan vallen en alleen in de praktijk kan worden gebracht als ze wordt gesteund door faculteitsbesturen en andere bestuurders, maar ook door de onderzoekers zelf.

Statistiekonderwijs zou studenten onder wie zich de latere onderzoekers bevinden, dus niet alleen moeten voorlichten over de diverse statistische methoden en de hiervoor beschikbare software, maar hen vooral moeten bijbrengen dat je niet mag verwachten dat je daarmee voldoende ervaring hebt om de goede beslissingen te nemen. De suggestie van velen (bijv. Asendorpf et al., 2013) om statistisch denken in het curriculum op te nemen is uitstekend, maar voorziet niet in de ervaring die pas ontstaat na jarenlange studie en onderzoek op het terrein van de statistiek zelf en het geven van statistische consultatie.

De bijdrage van methodologen is dat zij snel problemen herkennen die onderzoekers niet herkennen en dat ze dus voorzichtiger zijn dan leken. Dat lijkt een bescheiden claim, maar voorzichtigheid is bij de toepassing van een moeilijk vak als de statistiek een enorme winst. Als je in staat bent om op basis van een door ervaring gevoede intuïtie even gas terug te nemen en de tijd neemt om een probleem goed uit te zoeken, maak je veel minder fouten. Daarbij is het natuurlijk belangrijk dat je veel kennis hebt, want anders kom je niet echt verder. Maar kennis komt met ervaring, en dat geldt dus ook voor de juiste intuïtie.

OVER DE INTERPRETATIE VAN REPLICATIEONDERZOEK

Het onderzoek van Open Science Collaboration (2015) illustreert, naar ik vermoed, onbedoeld hoe moeilijk statistiek is, doordat men bij elk oorspronkelijk onderzoek

Statistiek gaat over het in kaart brengen van onzekerheid en dus over kansen

maar een replicatieonderzoek uitvoert. Het idee achter replicatieonderzoek is dat je vanwege steekproeftoevalligheden de resultaten van een onderzoek niet kunt vertrouwen. De vraag is dan of je de resultaten van twee onderzoeken wél kunt vertrouwen. Wat moet je concluderen als je de eerste keer een effect vindt en de tweede keer niet? En had je ook niet alle onderzoeken moeten repliceren die de eerste keer geen effect lieten zien? Die zijn echter door selection bias, het filedrawer effect en publication bias niet meer terug te vinden. En wat als je tweemaal een effect vindt? Volgt daaruit dat je de derde keer ook een effect zal vinden?

De interpretatie van replicatie-uitkomsten is dus lastig, en kan vanuit statistisch perspectief op verscheidene manieren worden beoordeeld – zie ook Open Science Collaboration (2015). Ik ga hier slechts kort op in en wil de lezer niet lastigvallen met techniek die meestal en terecht moeilijk wordt gevonden. Ik beperk me tot verdelingen van grotere aantallen replicatiestudies per onderzoek en wat zij betekenen voor de conclusies die je kunt trekken. Je kunt het aantal replicatiestudies van hetzelfde onderzoek inclusief het eerste onderzoek, samen m studies, opvatten als een steekproef uit alle mogelijke replicatiestudies van hetzelfde onderzoek. Dan levert 1 onderzoek plus 1 replicatieonderzoek een steekproef van m = 2 waarnemingen, en dat is wel heel weinig om conclusies uit te trekken. Stel nu dat iemand een groot aantal replicatiestudies heeft gedaan van een onderzoek waarin de nulhypothese waar is en elk significant resultaat dus ten onrechte de nulhypothese verwerpt; dit is een Type I-fout, en bij elke replicatiestudie loop je een kans op een Type I-fout die gelijk is aan het significantieniveau van de toets, vaak α=.05 . Die serie van replicatieonderzoeken lijkt wiskundig op een kansspel dat je m keer doet, elke keer met kans .05, en waarbij je kunt vragen hoe vaak je naar verwachting een Type I-fout maakt. Het rekenwerk doe je met behulp van de binomiale verdeling. Voor bijvoorbeeld zes replicatiestudies (m = 6) zijn de kansen op respectievelijk 0, 1, 2, 3, 4, 5, en 6 Type I-fouten gelijk zijn aan .7351, .2321, .0306, .0021 en driemaal .0000. De kans is dus bijna driekwart (.7351) dat je geen Type I fout maakt en ruim een kwart (.2649) dat je er minstens een maakt, zelfs als het onderzoek zich dus concentreert op een niet bestaand effect.

Ik besprak eerder dat veel onderzoek te gehaast plaatsvindt in kleine steekproeven waarin men enthousiast op zoek is naar effecten en negatieve resultaten negeert. In het voorbeeld met zes replicatiestudies betekent dit dus dat degene die dat ene significante resultaat vindt, mocht het zich toevallig voordoen, het wel publiceert, en dat dit dan alle aandacht zal krijgen (Ioannides, 2005; Nuijten, Van Assen, Veldkamp & Wicherts). Als er wel sprake is van een echt effect, zul je met veel grotere kans de nulhypothese verwerpen, wat dan terecht is. Maar het effect kan ook klein zijn en daardoor oninteressant, en het hangt van veel factoren af hoe je de resultaten precies moet interpreteren. Dit blijft een moeilijk onderwerp.

Daarom is mijn advies onderzoek te doen in grote steekproeven in plaats een relatief klein onderzoek vaker te herhalen. In grote steekproeven kun je effectgroottes erg precies schatten en veilig concluderen of het gevonden effect de moeite waard is.

CONCLUSIES

Mijn advies is bij de opzet van onderzoek en de statistische data-analyse een methodoloog in te schakelen. Verantwoord gebruik van statistiek vereist ervaring die onderzoekers meestal niet hebben. Gebrek aan ervaring bevordert het intuïtief kiezen van de verkeerde oplossingen voor methodologische en statistische problemen, die door hun moeilijkheid en tegen-intuïtieve karakter daartoe uitnodigen. Daarbij maken methodologen ook fouten, maar minder dan leken.

Over replicatieonderzoek heb ik maar weinig gezegd, en verwijs ik de lezer naar Bakker et al. (2012) voor de situatie waarin een effect echt is maar de power van de onderzoeksopzet te gering, en naar Van Assen, Van Aert & Wicherts (2015) voor de mogelijkheid om via meta-analyse conclusies te trekken op basis van een groot aantal verschillende onderzoeken naar hetzelfde fenomeen. Voor verdere discussie over replicatieonderzoek verwijs ik naar Pashler &

Wagenmakers (2012). Voor een uitvoeriger discussie die dit artikel mede inspireerde, verwijs ik naar Sijtsma (2016).

OVER DE AUTEUR

Klaas Sijtsma is hoogleraar Methoden van Psychologisch Onderzoek en decaan van de Tilburg School of Social and Behavioral Sciences. Adres: Tilburg School of Social and Behavioral Sciences, kamer P1.130 (Faculty Office), Tilburg University, Postbus 90153, 5000LE Tilburg. E-mail: k.sijtsma@tilburguniversity.edu. *De auteur dankt Andries van der Ark, Wilco Emons, Peter van der Heijden, Michèle Nuijten en Mark de Rooij voor hun commentaar op een eerdere versie van dit artikel.*

Summary

DOES PSCHOLOGY NEED REPLICATION STUDIES ?
K. SIJTSMA
A dependable interpretation of research results requires a sample of replication studies rather than just two, but psychological research is better off when effect sizes are estimated in a large sample. Prior to this conclusion, I argue that methodology and statistics is too difficult to practice on the side as most researchers do, thus leading to numerous derailments that tend to go unnoticed. Routinely soliciting expert advice from methodologists and statisticians will prevent many errors.

Literatuur

Abelson, R.P. (1995). *Statistics as principled argument*. Hillsdale, NJ: Elrbaum.
Asendorpf, J.B. et al. (2013). Recommendations for increasing replicability in psychology. *European Journal of Personality, 27*, 108-119.
Assen, M.A.L.M. van, Aert, R.C.M. & Wicherts, J.M. (2015). Meta-analysis using effect size distributions of only statistically significant studies. *Psychological Methods, 20*, 293-309.
Bakker, M., Van Dijk, A. & Wicherts, J. M. (2012). The rules of the game called psychological science. *Perspectives on Psychological Science, 7*, 543-554.
Busato, V.V. (2014). Van de geschiedenis leert men? Impact, publicatiedruk en fair play in de (psychologische) wetenschap. *Ons Erfdeel, 4*, 4-14.
Groot, A.D. de (1961). *Methodologie. Grondslagen van onderzoek en denken in de gedragswetenschappen*. 's Gravenhage: Mouton & Co.
Hand, D. (2014). *The improbability principle. Why coincidences, miracles and rare events happen every day*. London, UK: Penguin Random House.
Huff, D. (1954). *How to lie with statistics*. New York, London: Norton.
Ioannidis, J.P.A. (2005). Why most published research findings are false. *PLoS Med 2(8):e124*.
John, L.K., Loewenstein, G. & Prelec, D. (2012). Measuring the prevalence of questionable research practices with incentives for truth telling. *Psychological Science, 23*, 524-532.
Kadane, J.B., Schervish, M.J. & Seidenfeld, T. (1996). Reasoning to a foregone conclusion. *Journal of the American Statistical Association, 91*, 1228-1235.
Kahneman, D. (2011). *Thinking, fast and slow*. London: Penguin Books Ltd.
NRC Handelsblad (2015. Gedownload van http://www.nrc.nl/nieuws/2015/08/28/van-veel-psychologisch-onderzoek-blijft-weinig-over-als-je-het-herhaalt/
Nuijten, M.B., Van Assen, M.A.L.M., Veldkamp, C.L.S. & Wicherts, J.M. (2015). The replication paradox: Combining studies can decrease accuracy of effect size estimates. *Review of General Psychology, 19*, 172–182.
Open Science Collaboration (2015). Estimating the reproducibility of psychological science. *Science, 349 (6521)*, aac4716-1—aac4716-8.
Pashler, H. & Wagenmakers, E.J. (2012). Editors' introduction to the special section on replicability in psychological science: A crisis of confidence? *Perspectives on Psychological Science, 7*, 528-530.
Popper, K. (1935, 2002). *The logic of scientific discovery*. Abingdon, UK: Routledge.
Tversky, A. & Kahneman, D. (1974). Judgment under uncertainty: heuristics and biases. *Science, 185*, 1124-1131.
Ross, S.M. (1994). *A first course in probability* (fourth edition). New York: McMillan.
Sijtsma, K. (2016). Playing with data—Or how to discourage questionable research practices and stimulate researchers to do things right. *Psychometrika, 81*, 1-15.
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.
Volkskrant (2015). Gedownload van http://www.volkskrant.nl/binnenland/merendeel-psychologische-studies-lijkt-gebaseerd-op-drijfzand-a4130542/
Volkskrant (2015). Gedownload van http://www.volkskrant.nl/wetenschap/de-psychologie-van-gebrekkig-onderzoek-a4135765/