

# Schaarste is goed, maar de selectie moet beter

Lucas M. Seuren

Scarcity is good, but selection must improve

Hornikx & Batenburg (2016) demonstrate that publication pressure is an important reason why the social sciences face a problem with replicability. They argue that researchers result to questionable methods because the opportunities to publish are scarce. This paper, however, suggests that more room for publications will not lead to better research practices. Instead the scientific community should focus on stimulating good methods: (i) use power analyses to make sure an experiment is adequately sensitive and (ii) value the effect size over statistical significance.

Keywords: p-values; research integrity; inferential statistics; questionable research practices

## Inleiding

Het artikel van Hornikx & Batenburg is, zoals ze zelf ook beschrijven, er slechts één in een lange lijst van artikelen waarin problemen met onderzoeksintegriteit aan de kaak worden gesteld. Veel van deze artikelen komen uit de psychologie (bijvoorbeeld Asendorpf et al., 2013; Ferguson & Heene, 2012), maar het zou naïef zijn om aan te nemen dat de taalbeheersing, waar de methodiek vergelijkbaar is, niet met dezelfde problemen kampt. De bijdrage van Hornikx en Batenburg (2016) is daarmee een welkome eerste stap om te zien wat de integriteitsproblemen zijn binnen taalbeheersing en hoe die opgelost kunnen worden.

Het probleem zoals omschreven door Hornikx en Batenburg op basis van McQuarrie (2014) laat zien dat academici te maken hebben met een systeem dat leidt tot perverse prikkels. Hoewel opzettelijke fraude zoals in het geval van Diederik Stapel gelukkig zeldzaam is, blijkt uit een survey van John, Loewenstein en Prelec (2012) dat zogenaamde *questionable research practices*, zoals het afronden van *p*-waardes, aan de orde van de dag zijn. De ruime meerderheid van de 2.155 psychologen die de survey invulde gaf toe een van de bevraagde, onzuivere methoden te hebben toegepast. Erger nog: de respondenten vonden hun eigen gedrag veelal redelijk tot goed te rechtvaardigen. Alleen falsificatie van data werd sterk afgewezen.

De vraag die onderzoekers in de sociale wetenschappen al decennia bezighoudt is hoe we ervoor kunnen zorgen dat onzuivere methoden zoals in de survey van John et al. niet worden toegepast: dat onderzoek altijd integer wordt uitgevoerd. Het is een kwestie die recent weer veel aandacht krijgt, maar een die niet bepaald nieuw is (bijvoorbeeld Cohen, 1962; Rosenthal, 1979). Er zijn wel pogingen gedaan om bijvoorbeeld de *publicatiebias* te doorbreken door het publicatiebeleid van tijdschriften te wijzigen, maar de meeste strandden al binnen enkele jaren (Giner-Sorolla, 2012): zo werd in 1979 *Replications in Social Psychology* opgericht, maar er verschenen maar drie delen.

Aangezien er al lange tijd aandacht is voor de problemen binnen kwantitatief onderzoek, lijkt het gebrek aan verandering niet het gevolg van een gebrek aan bewustzijn bij de onderzoeksgemeenschap. De publicatiebias en matige repliceerbaarheid van gepubliceerd onderzoek worden breed uitgemeten in menig artikel en er zijn zelfs special issues aan gewijd, zoals in het

*European Journal of Personality* (Asendorpf et al., 2013). Waarom blijven structurele hervormingen dan veelal uit, en stranden de weinige pogingen in een vroeg stadium?

In dit artikel bespreek ik één van de in mijn ogen fundamentele problemen: gebrekkige statistische kennis leidt tot een gebrekkig peer reviewproces, waardoor het loont om onderzoek niet integer uit te voeren. Maar voor ik inga op het reviewproces, laat ik zien dat de schaarste van publicatieruimte in wetenschappelijke tijdschriften niet de oorzaak is van het probleem. Schaarste is juist in het belang is van de wetenschap. Daarna beargumenteer ik dat het daadwerkelijke probleem is dat de huidige selectiecriteria van tijdschriften niet adequaat zijn: door gebrekkige kennis van statistiek worden zogenaamd statistisch significante uitkomsten overgewaardeerd. Schaarste speelt een rol bij fraude, maar een inadequaat peer reviewproces is de katalysator voor onzuivere onderzoeksmethoden.

## Schaarste is in het belang van onderzoekers

Zoals terecht wordt opgemerkt door Hornikx en Batenburg, worstelen veel onderzoekers met de schaarste van publicatieruimte. Maar zoals ze ook laten doorschemeren: deze schaarste is kunstmatig en daarmee eenvoudig te verhelpen. Het internet biedt potentieel oneindig veel ruimte voor tijdschriften om artikelen te produceren, met slechts geringe extra kosten voor de grotere datastromen (McQuarrie, 2014). De vraag is of dat wenselijk is.

Er zijn twee redenen om publicatieruimte schaars te houden. De eerste is dat als er meer artikelen ingediend worden, de werkdruk voor reviewers en editors beduidend hoger wordt. Onderzoekers moeten een groter deel van hun tijd gaan besteden aan reviewwerk; tijd die ze ongetwijfeld liever in hun eigen onderzoek steken. Ten tweede zijn meer artikelen, zoals ik hieronder uitwerk, niet in het belang van individuele academici en daarmee ook niet voor wetenschappelijk onderzoek in bredere zin.

Het doel van de wetenschap is het vergroten van de menselijke kennis, dat is althans het ideaal. Individuele onderzoekers hebben meer belangen. Onderzoeksresultaten moeten gedeeld worden met de onderzoeksgemeenschap, en moeten dus opgeschreven, gepubliceerd, en gelezen worden. En daar blijft het niet bij. De kennis die we nastreven moet ook enig nut hebben. Dat nut meten we voornamelijk af aan twee aspecten: (i) het prestige van het tijdschrift waarin onderzoeksresultaten gepubliceerd worden – we moeten zoveel mogelijk publiceren in het hoogste kwartiel van de tijdschriften, liefst zelfs in de beste 10% – en (ii) de frequentie waarmee de resultaten door andere onderzoekers geciteerd worden. Deze drijfveer om veel te publiceren en veel geciteerd te worden staat breder bekend als *Publish or Perish*.

Door meer ruimte te creëren voor artikelen wordt de publicatiedrang dus niet opgelost. Onderzoek zal nog altijd gepubliceerd, gelezen, en geciteerd moeten worden. Sterker nog, het verergert een bestaand probleem: nu al worden veel artikelen niet geciteerd en soms zelfs niet eens gelezen. Door meer te publiceren zal het relatieve aantal ongeciteerde artikelen toenemen; de tijd die onderzoekers hebben om die artikelen te lezen is nu eenmaal eindig.

Doordat artikelen relatief minder gelezen worden, zal het gemiddeld aantal citaties per artikel, en daarmee de impactfactor van een tijdschrift, ook sterk dalen. Dat is nadelig voor tijdschriften, want een hoge impactfactor betekent dat onderzoekers hun beste werk naar die tijdschriften sturen. Voor de onderzoekers zelf is die hoge impactfactor ook van belang. De kwaliteit van onderzoekers wordt in belangrijke mate afgelezen aan de impact van zijn/haar onderzoek.

Zelfs al zou al het onderzoek gepubliceerd kunnen worden, dan lopen we dus alleen maar tegen nieuwe problemen aan. Bovendien lossen we het werkelijke probleem niet op. Zolang onderzoekers met elkaar moeten concurreren voor banen en onderzoeksgeld, zullen selectiecommissies overzichtelijke kwaliteitscriteria moeten hanteren. En zolang er kwaliteitscriteria

zijn, zullen er onderzoekers zijn die proberen om op minder dan zuivere methoden aan die criteria te voldoen.

Het doel van de onderzoeksgemeenschap moet niet zijn om zoveel mogelijk te publiceren, maar ervoor te zorgen dat gepubliceerd onderzoek voldoet aan methodologische standaarden. Het wapen dat academisch onderzoek altijd heeft gehad tegen onzuivere onderzoeksmethoden is het peer reviewproces, en daar gaat het voor een groot deel mis. Want hoewel peer review zeker geen zaligmakend middel tegen fraude is, zoals McQuarrie (2014) ook opmerkt, werkt het zeker niet als er slechte kwaliteitscriteria gehanteerd worden waarin perverse prikkels gestimuleerd worden.

## Kennis is de basis

In experimenteel onderzoek wordt er veel waarde gehecht aan statistisch significante resultaten. Veel van de door John et al. (2012) aangehaalde twijfelachtige praktijken zijn ook gericht op publiceerbare  $p$ -waarden. Hornikx en Batenburg zoeken de oplossing onder andere in de verslaglegging door onderzoekers. Zo zouden tijdschriften de statistische toetsing moeten controleren zodat “de gebruiker kan zien in hoeverre de gerapporteerde  $p$ -waarden kloppen.” (2016, [pagina volgt bij drukproef] Maar de waarde van een significant resultaat, het nut van de  $p$ -waarde, wordt daarmee overschat. Hieronder werk ik een aantal redenen uit: (1) het gaat voorbij aan de gevoeligheid, i.e., de statistische power, van een experiment; (2) het wetenschappelijk nut van onderzoeksresultaten is minder afhankelijk van het daadwerkelijk gemeten effect; en (3) het geeft de illusie van precisie en geeft geen antwoord op de vraag die de onderzoekers bij hun experiment stellen.

Dat statistische power in een experiment van belang is, zal bij elke onderzoeker wel bekend zijn. Toch wordt de waarde van statistische power onderschat, en daarom regelmatig niet genoemd bij onderzoeksresultaten (Kühberger, Fritz, & Scherndl, 2014). Power geeft niet alleen aan hoe gevoelig een experiment is, op de lange termijn biedt het inzicht in hoeveel van de gevonden significante resultaten er daadwerkelijk een effect is. Colquhoun (2014) laat zien dat het aantal valse positieven (gepubliceerde type I-fouten) ten opzichte van het totaal aantal positieven ook afhangt van de power van experimenten. Doordat de power van veel experimenten laag is – in de psychologie ligt de gemiddelde power tussen de 0,35 en 0,65 (Asendorpf et al., 2013) – ligt het aantal valse positieven veel hoger dan 5%. Dit percentage kan deels worden ingeperkt door betere hypothesen te formuleren, dat wil zeggen, hypothesen waarvan de verwachting dat ze kloppen hoog is. Maar als we alleen hypothesen testen waarvan we weten dat ze kloppen hoeven we ook geen experimenten te doen. Het is dus noodzakelijk dat experimenten voldoende power hebben. Elk experiment zou dan ook voorafgegaan moeten worden door een power-analyse en de resultaten daarvan zouden genoemd moeten worden in elke publicatie.

Het tweede probleem van de preoccupatie met de  $p$ -waarde is dat niet elk effect betekenisvol is. We willen niet alleen weten of een effect bestaat, maar ook hoe groot het is. Dit hangt deels samen met power, aangezien verwachte effectgrootte een van de invoerwaarden is voor een power-analyse. Maar daarnaast maakt effectgrootte het veelal gemakkelijker om de  $p$ -waarde te interpreteren. Zo is een significant effect met een zeer klein tot verwaarloosbaar effect (e.g., Kramer, Guillory, & Hancock, 2014 over *emotional contagion* op Facebook) mogelijk volstrekt oninteressant. Door effectgrootte standaard te vermelden wordt ook voorkomen dat er een beeld ontstaat dat “de mate van significantie” inzicht geeft in hoe groot een effect is (Kühberger, Fritz, & Scherndl, 2014).

Het derde, onderliggende probleem is dat het concept  $p$ -waarde door de overgrote meerderheid van onderzoekers, zowel beginnend als ervaren, verkeerd begrepen wordt (Gigerenzer, 2004). Bij nulhypothese-toetsend onderzoek is het doel om een uitspraak te kunnen doen over de

waarschijnlijkheid van de (alternatieve) hypothese. Maar de methoden die worden toegepast bieden dat niet. De  $p$ -waarde geeft de kans dat we de gevonden data zouden vinden ervan uitgaande dat de nulhypothese waar is. Zelfs al is die kans klein, daarmee kunnen we nog geen uitspraken doen over de waarschijnlijkheid van de nulhypothese. Met een kleine kans lijkt het aannemelijk dat de nulhypothese onwaar is, maar die correlatie is zwak (Trafimow & Rice, 2009).

Helaas is een goed alternatief niet voorhanden. Bayesiaanse statistiek geeft weliswaar de waarschijnlijkheid van de nulhypothese op basis van de gevonden data, maar daarvoor moeten aannames gedaan worden die lang niet altijd waar zijn (Trafimow, 2005).

De beste oplossing lijkt te zijn om de problemen van nulhypotheseonderzoek te onderkennen en het criterium van statistische significantie los te laten. Het leidt immers alleen maar tot, veelal frauduleuze, pogingen om aan een slecht criterium te voldoen. Met andere woorden, we moeten niet langer een onderscheid maken tussen significante en niet-significante resultaten. Dat wil niet zeggen dat alles gepubliceerd kan worden: we moeten alleen andere criteria hanteren. Zo moet het experiment gevoelig genoeg zijn, zodat een voorspeld effect gevonden kan worden en repliceerbaar is. Daarnaast moet de grootte van het gemeten effect helder zijn. Goed onderzoek geeft altijd interessante resultaten, daar hebben we geen criterium als de  $p$ -waarde voor nodig.

## Referenties

- Asendorpf, J.B., Conner, M., De Fruyt, F., De Houwer, J., Denissen, J.J.A., Fiedler, K., ... Wicherts, J.M. (2013). Recommendations for increasing replicability in psychology. *European Journal of Personality*, 27, 108-119.
- Cohen, J. (1962). Statistical power of abnormal-social psychological research. *Journal of Abnormal and Social Psychology*, 65, 145-153.
- Colquhoun, D. (2014). An investigation of the false discovery rate and the misinterpretation of  $p$ -values. *Royal Society Open Science*, 1, 140216.
- Ferguson, C.J., & Heene, M. (2012). A vast graveyard of undead theories: Publication bias and psychological science's aversion to the null. *Perspectives on Psychological Science*, 7(6), 555-561.
- Gigerenzer, G. (2004). Mindless statistics. *The Journal of Socio-Economics*, 33, 587-606.
- Giner-Sorolla, R. (2012). Science or art? How aesthetic standards grease the way through the publication bottleneck but undermine science. *Perspectives on Psychological Science*, 7, 562-571.
- Hornikx, J. & Batenburg A. (2016). Integriteit in kwantitatief, empirisch onderzoek: problemen en mogelijke oplossingen. *Tijdschrift voor Taalbeheersing*, 38
- John, L.K., Loewenstein, G., & Prelec, D. (2012). Measuring the prevalence of questionable research practices with incentives for truth telling. *Psychological Science*, 23(5), 524-532.
- Kramer, A.D.I., Guillory, J.E., Hancock, J.T. (2013). Experimental evidence of massive-scale emotional contagion through social networks. *Proceedings of the National Academy of Sciences*, 111(24), 8788-8790.
- Kühberger, A., Fritz, A., & Scherndl, T. (2014). Publication bias in psychology: A diagnosis based on the correlation between effect size and sample size. *PLOS ONE*, 9(9). DOI: <http://dx.doi.org/10.1371/journal.pone.0105825>
- McQuarrie, E.F. (2014). Threats to the scientific status of experimental consumer psychology: A Darwinian perspective. *Marketing Theory*, 14, 477-494.
- Rosenthal, R. (1979). The file drawer problem and tolerance for null results. *Psychological Bulletin*, 86, 638-641.

Trafimow, D. (2005). The ubiquitous Laplacian assumption: reply to Lee and Wagenmakers.  
*Psychological Review*, 112, 669-674.

Trafimow, D. & Rice, S. (2009). A test of the null hypothesis significance testing procedure correlation argument. *The Journal of General Psychology*, 136, 261-269.

Over de auteur

Lucas M. Seuren is promovendus aan de Rijksuniversiteit Groningen binnen het Center for Language & Cognition. E-mail: l.m.seuren@rug.nl.