

## Psychologisch onderzoek

# Te grote stelligheid bedreigt kwaliteit wetenschap

Rink Hoekstra

**In psychologisch onderzoek wordt om praktische redenen veel gebruik gemaakt van steekproeven. Voor het generaliseren van deze steekproefgegevens is statistiek nodig. Het blijkt dat onderzoekers deze statistische technieken relatief slecht beheersen. Met name de significantietoets, die in bijna ieder wetenschappelijk artikel wordt gebruikt, blijkt aanleiding te geven tot veel misinterpretaties. Deze misinterpretaties zijn vaak ernstig, en zouden zelfs kunnen leiden tot een wetenschap waarin relatief veel onzin als zinvol wordt verkocht.**

Onderzoekers in de psychologie bestuderen het menselijk gedrag en de onderliggende mentale processen. Soms worden individuen bestudeerd, en eventuele conclusies zullen dan meestal vooral op dat individu van toepassing zijn. Dit is bijvoorbeeld het geval bij een zogenaamde  $n=1$ -studie. Vaak ook wil je als psycholoog een uitspraak doen over een grotere groep mensen, of misschien zelfs over de meeste mensen. Dit is niet altijd eenvoudig, en wel om twee redenen: ten eerste verschillen mensen nogal van elkaar, en moet je dus uiterst voorzichtig zijn met het doen van algemene uitspraken. Ten tweede zijn er praktische gronden waarom het in de meeste gevallen onmogelijk is alle personen van de groep waarin je als onderzoeker geïnteresseerd bent te gebruiken voor je onderzoek. Dit kost namelijk vaak te veel tijd en/of geld, of is praktisch bijkans onuitvoerbaar. Als je bijvoorbeeld wilt onderzoeken of mannen en vrouwen verschillen op een bepaald punt kun je moeilijk alle mannen en vrouwen op de wereld gaan onderzoeken.

Om dit laatste probleem van een te grote onderzoeksgroep op te lossen wordt vaak gebruik gemaakt van steekproefgebaseerd onderzoek. Begonnen wordt met een willekeurige steekproef uit de groep waarover je iets wilt weten. Die groep wordt ook wel populatie genoemd. Datgene wat je wilt weten over die populatie bereken je voor de mensen in je steekproef. Zogenaamde inferentiële statistiek is vervolgens nodig om de gevonden uitkomsten te generaliseren naar de populatie. Dat wil zeggen: je wilt op basis van je steekproefresultaat schatten hoe de situatie is in de populatie, en daarbij ook rekening houden met toeval. Helaas lijkt hierbij in de praktijk nogal eens wat fout te gaan.

De meest gebruikte generalisatietechniek binnen de sociale wetenschappen (maar waarschijnlijk ook binnen veel andere vakgebieden) is de zogenaamde significantietoets. Deze toets wordt zeker in psychologisch onderzoek bijna reflexmatig gebruikt: weinigen staan er bij stil dat er überhaupt alternatie-

ven voor deze toets bestaan. Uit verschillende onderzoeken blijkt dat in wetenschappelijke tijdschriften over psychologie minstens 90% van alle artikelen één of meer significantietoetsen bevat. Ook in wetenschapskaternen van dagbladen of in populair wetenschappelijke literatuur wordt er kwistig gestrooid met de term significant. Men zou dus verwachten dat deze bijna altijd gebruikte toets voor onderzoekers dusdanig bekend is dat er niet of nauwelijks fouten mee zullen worden gemaakt, en dat als ze al gemaakt worden, ze door de rest van de wetenschappelijke omgeving zullen worden gecorrigeerd. Ook zou je kunnen verwachten dat de bruikbaarheid van een zo frequent gebruikte toets nauwelijks ter discussie staat. Het blijkt echter toch iets anders te liggen.

### Achtergronden van de significantietoets

Bij het gebruik van een significantietoets start je met een hypothese (meestal nulhypothese genoemd) waarvan je wilt aantonen dat deze niet waar is. Op deze manier probeer je aannemelijk te maken dat datgene wat je verwacht, -de zogenaamde alternatieve hypothese-, wél waar is. Zo zou je, als je wilt aantonen dat een bepaalde behandeling beter werkt dan een eerdere behandeling, als nulhypothese kunnen aannemen dat de nieuwe behandeling net zo goed werkt als de eerdere. Vervolgens probeer je bewijs te verzamelen tegen deze nulhypothese, en dit doe je door je steekproefuitkomst te vergelijken met deze hypothese. Dit gaat als volgt:

- Het bepalen van de kans: Je berekent de kans dat je de uitkomst zou vinden die je gevonden hebt of extremer, onder de aanname dat de nulhypothese waar zou zijn (waarvan je dus hoopt aan te kunnen tonen dat dit niet het geval is). Deze kans wordt  $p$ -waarde genoemd.
- Het bepalen of die kans significant is: Is deze kans kleiner dan een vooraf bepaalde grenswaarde (bijvoorbeeld 0.05),

dan noemen we het effect significant, en ‘verwerpen we de nulhypothese’.

- Het trekken van een conclusie: Bij een significant effect concluderen we dat er ‘waarschijnlijk ook in de populatie een effect is’. Fout is in dit geval dus te stellen dat er zeker een effect in de populatie is. Bij een niet-significant effect kunnen we eigenlijk niet zo veel concluderen, hoogstens dat je steekproefeffect blijkbaar niet groot genoeg was om de nulhypothese te verwerpen. Een niet-significant resultaat betekent dus niet dat je mag concluderen dat er waarschijnlijk geen effect is (ook al blijkt deze verleiding soms groot)!

Bovenstaande procedure kan misschien iets duidelijker gemaakt worden aan de hand van een voorbeeld. Een psycholoog wil onderzoeken of spinnenangst meer afneemt door behandeling A dan de al bekende behandeling B. Om dit te meten wijst ze mensen uit een willekeurige steekproef van mensen met een hoge mate van spinnenangst aan één van beide behandelingen toe. Van iedere persoon wordt de angst voor en na de behandeling gemeten. Laten we aannemen dat behandeling A in de steekproef inderdaad gemiddeld genomen tot een grotere afname van spinnenangst heeft geleid dan behandeling B. Nu kan de kans uitgerekend worden (stap 1) om minstens zo’n gemiddeld verschil in afname van spinnenangst te vinden, wanneer zou gelden dat behandeling A in werkelijkheid net zo goed zou werken als behandeling B. Als in dit voorbeeld die kans 0.01 is en als grenswaarde 0.05 wordt gebruikt betekent dit dat we een significant effect gevonden hebben (stap 2). We kunnen nu concluderen dat onze behandeling A in de populatie waarschijnlijk inderdaad gemiddeld beter werkt dan behandeling B (stap 3), waarmee uiteraard niet gezegd is dat dit voor iedere patiënt het geval is.

Samenvattend probeer je dus bij een significantietoets aan te tonen dat het onwaarschijnlijk is het gevonden steekproefeffect te vinden als de nulhypothese waar zou zijn. Bij een significant effect (en dus een relatief kleine p-waarde) zijn er twee mogelijkheden: of je hebt toevallig in deze steekproef een heel extreem effect gevonden, terwijl er in werkelijkheid in de populatie helemaal geen verschil is, of die nulhypothese is helemaal niet waar en er is wel degelijk een verschil in de populatie. Gezien het feit dat de eerste verklaring wel erg toevallig is, ‘kies’ je er bij een significante uitkomst voor de tweede verklaring aannemelijker te achten. Je neemt daarbij op de koop toe dat in werkelijkheid misschien de eerste verklaring juist zou kunnen zijn. Dit is niet helemaal onlogisch en ook wel enigszins intuïtief: als je 10 keer met een nieuwe dobbelsteen gooit en die komt 10 keer op ‘zes’, denk je waarschijnlijk dat er sprake is van een valse dobbelsteen, ook al is het natuurlijk theoretisch ook mogelijk dat je, heel toevallig, 10 keer een zes met een eerlijke dobbelsteen hebt gegooid.

### De significantietoets door de jaren heen

Historisch gezien is de rol van de significantietoets minder prominent geweest dan je op basis van het gebruik nu misschien

zou verwachten. Als je 1879 (het jaar waarin Wilhelm Wundt zijn psychologisch laboratorium in Leipzig stichtte) als de geboorte neemt van de psychologie als wetenschap, dan speelde die toets in de eerste helft van haar bestaan tot nu toe nauwelijks een rol. Hoewel je zou kunnen stellen dat de significantietoets (of iets wat daar sterk op lijkt) eind jaren ’20 is bedacht, nam het gebruik ervan pas rond het midden van de vorige eeuw een hoge vlucht. Vanaf ongeveer de zeventiger jaren lijkt in de praktijk van steekproefgebaseerd onderzoek de significantietoets nauwelijks meer weg te denken. Dit betekent echter niet dat de significantietoets niet werd bekritiseerd. In 1938 schreef Joseph Berkson al dat de techniek niet goed bruikbaar zou zijn, en sindsdien is de hoeveelheid kritiek bijna letterlijk exponentieel gestegen, met alleen al in de jaren ’90 een kleine 200 artikelen waarin de significantietoets wordt bekritiseerd (Kline, 2004). Uiteindelijk heeft deze kritiek de beroepsvereniging van psychologen American Psychological Association (APA) eind vorige eeuw doen besluiten een adviescommissie in te stellen om de problematiek rond de significantietoets te onderzoeken. De suggesties van deze zogenaamde Task Force on Statistical Inference (Wilkinson, 1999) zijn uiteindelijk deels verwerkt in de laatste twee edities van de APA-manual (American Psychological Association, 2001 en 2009). Deze handleiding wordt door vele onderzoekers binnen en buiten de psychologie als richtinggevend voor het formuleren van wetenschappelijke publicaties beschouwd.

### De significantietoets onder vuur

Zoals gezegd is er dus al decennialang veel kritiek geuit op de significantietoets. De belangrijkste problemen met de significantietoets zijn de volgende (deels na te lezen in het goed leesbare ‘The earth is round ( $p < .05$ )’ van Jacob Cohen, 1994):

#### **Waarschijnlijk is in de praktijk de nulhypothese (bijna) nooit waar**

Het lijkt bijvoorbeeld moeilijk een variabele te vinden waarop mannen en vrouwen gemiddeld genomen precies hetzelfde scoren (tot op elke decimaal achter de komma), en het lijkt lastig voor te stellen dat twee variabelen precies 0 correleren. Waarom zou je proberen een hypothese te verwerpen, als je eigenlijk van te voren toch al weet dat deze eigenlijk verworpen zou moeten worden? Deze stelling maakt het uitvoeren van een significantietoets tot een vrij zinloze exercitie: feitelijk probeer je een vraag beantwoord te krijgen waarop je het antwoord toch al weet.

#### **De significantietoets beantwoordt niet de vraag die je als onderzoeker beantwoord zou willen krijgen**

Als je al wilt weten of er een effect is in de populatie geeft de significantietoets je niet rechtstreeks antwoord op die vraag. Als onderzoeker zou je eigenlijk willen weten: wat is de kans, gegeven mijn steekproefuitkomst, dat er inderdaad een effect in de populatie is (al valt gezien het eerste probleem ook te bezien of dit de meest interessante vraag is). De significantietoets geeft echter de kans op het gevonden effect of extremer, als

de nulhypothese waar zou zijn. Dit lijkt misschien erg sterk op elkaar, maar dit is zeker niet hetzelfde.

Kortom: de significantietoets geeft antwoord op een niet-gestelde vraag: alsof je met een hamer een schroef in de muur probeert te slaan.

### **De significantietoets moedigt mensen aan dichotoom ('zwart-wit') te denken**

Dit is misschien wel het ernstigste probleem van de significantietoets. Door wetenschappelijke uitkomsten te categoriseren in significante resultaten en niet-significante resultaten vindt een grove versimpeling van de werkelijkheid plaats, die het risico met zich meebrengt dat dit ook als een scheidslijn tussen belangrijk en niet zo belangrijk onderzoek wordt gezien. De betekenis van het Engelse woord significant (letterlijk: 'betekenisvol') draagt hier zeker aan bij. De werkelijkheid is echter veel te complex om te beschrijven in zwart-wit termen. Een risico is dat hierdoor het al dan niet significant zijn centraal komt te staan, terwijl relatief weinig aandacht uitgaat naar de grootte van een effect, die vaak veel interessanter is. Een heel klein maar oninteressant effect kan significant zijn, terwijl een groot en interessant effect niet-significant kan zijn.

### **Door de significantietoets wordt een zogenaamd 'file drawer'-probleem gecreëerd.**

Dit probleem ligt redelijk in het verlengde van het vorige probleem. Als mensen niet-significante resultaten zien als niet belangrijke resultaten, is het wel denkbaar dat auteurs die een niet-significant effect hebben gevonden niet eens de moeite nemen dit artikel bij een wetenschappelijk tijdschrift in te dienen. Ook is het goed voor te stellen dat het artikel als het al wordt ingediend vanwege het niet-significante effect wordt geweigerd. De onderzoeken verdwijnen dus ongepubliceerd in een la ('file drawer'). Wanneer zou gelden, -en er is geen reden om aan te nemen van niet-, dat het onderzoek achter die niet-significante resultaten even goed is uitgevoerd als onderzoek met wél significante resultaten, dan vindt er in de gepubliceerde onderzoeken dus een structurele vertekening van de werkelijkheid plaats omdat een deel van de onderzoeken wordt achtergehouden (bijvoorbeeld Rosenthal, 1979).

### **Alternatieven voor de significantietoets**

Doordat er bijna standaard een significantietoets gebruikt wordt zou je bijna vergeten dat er wel degelijk alternatieven of aanvullingen zijn voor de significantietoets. Ik zal hieronder twee beschrijven. Dit is echter wel een selectie: de in deze context vaak gepropageerde Bayesiaanse statistiek laat ik hier bijvoorbeeld buiten beschouwing, omdat deze relatief complex is om uit te leggen aan mensen die hier nog niet eerder mee in aanraking gekomen zijn.

#### **Meer nadruk op effectgrootte**

Zoals eerder gezegd is een risico van het gebruik van significantietoetsen dat deze mensen aanspoort relatief zwart-wit over hun uitkomsten te denken. Er wordt waarschijnlijk eerder gekeken naar of er überhaupt een effect is of niet (waarbij zoals eerder

beweerd de afwezigheid van een effect niet eens aannemelijk gemaakt kan worden) en daardoor minder naar de grootte van een effect. Dit zou veel meer centraal moeten komen te staan. Een niet-significant groot effect is misschien wel veel interessanter dan een minuscuul maar significant klein effectje. Let wel: niet ieder groot effect is per se interessant, maar zeker zo sterk geldt dat niet ieder significant effect interessant is. Binnen de statistiek is er bij mijn weten niemand te vinden die bovenstaand punt bestrijdt: volgens de eerdergenoemde Taskforce (Wilkinson, 1999) is het rapporteren en interpreteren van effectgrootte zelfs essentieel voor het doen van goed onderzoek.

### **Betrouwbaarheidsintervallen**

Een belangrijk alternatief voor de significantietoets is het betrouwbaarheidsinterval (bijvoorbeeld Thompson, 2007). Bij deze techniek worden eigenlijk de informatie uit de significantietoets en effectgrootte samengenomen, en in een interval uitgedrukt. Vaak wordt een 95%-betrouwbaarheidsinterval gebruikt, waarbij gegeven is dat normaalgesproken in 95% van de gevallen de populatiewaarde waar je naar op zoek bent binnen het berekende interval ligt. Om een voorbeeld te noemen: stel dat je voor het eerdergenoemde spinnenangst-voorbeeld vindt dat therapie A voor een afname zorgt van 15 punten (op bijvoorbeeld een 100-punts-schaal), en stel dat het bijbehorende 95%-betrouwbaarheidsinterval loopt van 10 tot 20. Dit betekent dat je (uiteraard) niet precies weet hoe veel de afname gemiddeld genomen in de populatie zou zijn, maar dat het waarschijnlijk ergens in de range van 10 tot 20 ligt. Hierdoor beantwoord je niet alleen de vraag of de therapie werkt, maar tegelijkertijd ook hoe sterk die werking ongeveer is. Dit geldt algemeen: onderzoeken die vragen naar de grootte van een effect beantwoorden, beantwoorden tegelijkertijd de vraag of er een effect is, en zijn daarom bijna altijd te prefereren. Onder andere om die reden wordt in de laatste twee edities van de APA-manual het gebruik van betrouwbaarheidsintervallen dan ook 'sterk aangeraden'.

### **Hoe zit het in de praktijk: de uitkomsten van mijn proefschrift**

Er is dus in de afgelopen jaren ontzettend veel kritiek geweest op nu net die toets die bij het generaliseren van steekproefdata verreweg het meest gebruikt wordt. Deze bezwaren waren echter zelden gebaseerd op onderzoek naar het gebruik van die toets: meestal ging het om statistici of wetenschapsfilosofen die fundamentele kritiek hadden op de toets als zodanig. Er was dus relatief weinig bekend over hoe die toets nu in de praktijk gebruikt wordt. Zijn onderzoekers zich eigenlijk bewust van de nadelen of problemen van de significantietoets? Als dat het geval zou zijn is het hierboven beschreven probleem namelijk misschien wel kleiner dan door de critici wordt beweerd. Wordt de toets in de praktijk met enige voorzichtigheid geïnterpreteerd, of worden er juist vaak ongenueanceerde en dichotome conclusies getrokken? In mijn proefschrift *Use and Usability of Inferential Techniques* (Het Gebruik en de Bruikbaarheid van Inferentiële Technieken, 2009) heb ik onderzocht hoe steekproefgegevens in de praktijk

worden gegeneraliseerd. Hoe wordt de significantietoets gebruikt, en wat voor conclusies worden hier aan verbonden? Worden alternatieven, zoals het betrouwbaarheidsinterval, gebruikt, en indien dat het geval is: worden deze op de juiste manier geïnterpreteerd? In verschillende onderzoeken heb ik geprobeerd meer zicht te krijgen op deze praktijk. De resultaten geven een redelijk schokkend beeld van de manier waarop onderzoekers hun steekproefdata analyseren en hierover conclusies trekken.

### **De praktijk in gepubliceerde artikelen**

Het belangrijkste communicatiemiddel voor wetenschappers is het publiceren in (vaak internationale) tijdschriften. Over het algemeen sturen onderzoekers hun manuscripten naar de tijdschriftredacties, en die laten vervolgens andere experts oordelen over de kwaliteit van het artikel, waarna de redactie besluit of het artikel in al dan niet aangepaste vorm geplaatst kan worden. Deze selectiemethode zou moeten zorgen voor een redelijk niveau van de te publiceren manuscripten.

Voor het bestuderen van de praktijk van het generaliseren en analyseren van data heb ik een kleine 300 artikelen van *Psychonomic Bulletin & Review* doorgenomen, een gerenomeerd wetenschappelijk tijdschrift waarin allerhande psychologische onderwerpen aan bod komen. Zoals verwacht was de significantietoets verreweg de meest gebruikte techniek (deze toets kwam in 97% van de artikelen minstens eenmaal voor), ondanks het feit dat dit niet expliciet door het tijdschrift noch door de APA-manual werd geëist. Betrouwbaarheidsintervallen echter, die in de APA-manual sterk worden aangeraden, werden bijna nooit vermeld (in slechts 7% van de gevallen).

Twee belangrijke interpretatiefouten kwamen veelvuldig voor: de ernstige fout van het aannemen van de nulhypothese, waarbij dus een niet-significant resultaat beschouwd wordt als bewijs voor de afwezigheid van een effect was te vinden in meer dan de helft van de artikelen. Ook het met stelligheid verwerpen van de nulhypothese op basis van een significant resultaat ('het effect is significant, dus ik ben zeker dat er een populatie-effect bestaat') kwam regelmatig voor (in zo'n 20% van de artikelen). Beide uitkomsten ondersteunen het idee dat de significantietoets dichotoom wordt geïnterpreteerd. Effectgroottes (vooral gemiddeldes) werden in bijna alle artikelen vermeld. Deze werden echter bijna nooit geïnterpreteerd, waardoor de indruk leek te worden gewekt dat dit voor de auteurs niet erg van belang was (terwijl dit misschien juist het interessantste deel is).

### **De praktijk op de werkvloer**

Een mogelijke verklaring voor het in de vorige paragraaf beschreven gedrag zou kunnen zijn dat onderzoekers eigenlijk wel weten hoe het zou moeten, maar zich aanpassen aan de heersende cultuur waarin het zoals beschreven vaak fout gaat. Dit is niet heel gek: als je als nieuwkomer een positie probeert te veroveren in de wetenschappelijke wereld ligt het voor de hand je aan te passen aan de in die wereld geldende gewoontes, ook al zou je het zelf misschien anders willen doen. Om die reden heb

ik een dertigtal promovendi bestudeerd tijdens het analyseren van fictieve data op hun eigen werkplek.

Ook in dit onderzoek werd echter het eerdere beeld bevestigd: de promovendi leken geneigd hun analyses vrij 'quick and dirty' uit te voeren, ondanks de instructie hun analyses zo uit te voeren als ze anders ook zouden doen en ondanks dat ze zich bewust waren van het feit dat ze geobserveerd werden. Zo werden er zelden plaatjes van de gegevens gemaakt (wat toch een goede en misschien wel onmisbare is om inzicht in je gegevens te krijgen), en werd er zelden nagegaan of de data aan de gewenste voorwaarden voor de uit te voeren techniek voldeden (zoals bijvoorbeeld het nagaan of er sprake was van een normale verdeling, wat voor sommige technieken een vereiste is). Wanneer hun gevraagd werd de conclusies te beschrijven (waarbij niet gezegd werd hoe zij dit moesten doen) bleken alle proefpersonen hiervoor de significantietoets te gebruiken, en bleek de interpretatie erg zwart-wit. Betrouwbaarheidsintervallen, ondanks dat die dus al jaren sterk worden aangeraden in de belangrijkste handleiding voor het doen van onderzoek, werden überhaupt niet opgesteld.

### **Wat te doen? Tips voor onderzoekers en lezers van wetenschappelijke artikelen**

Bovenstaande resultaten beschrijven een wetenschappelijke omgeving waarin de significantietoets in de afgelopen 80 jaar een centrale plek heeft gekregen bij het generaliseren van steekproefgegevens. Weinig onderzoekers lijken zich echter bewust te zijn van de nadelen van deze toets en van de beschikbaarheid van alternatieve methodes. De critici die, zoals eerder beschreven, wezen op het gevaar van de significantietoets worden door mijn onderzoek dus ook met resultaten uit de wetenschappelijke praktijk ondersteund: aan de significantietoets worden inderdaad vaak incorrecte conclusies verbonden, en lijkt onderzoekers er inderdaad toe aan te zetten om hun data dichotoom te interpreteren. Hoewel er statistici zijn die wél een toekomst zien voor de significantietoets lijkt er bij de meerderheid consensus te bestaan over het feit dat dergelijke dichotome uitspraken hoe dan ook onwenselijk zijn.

Voor wetenschappers of voor mensen die beroepshalve regelmatig wetenschappelijke artikelen lezen lijkt dit een vrij deprimerende boodschap: als de meeste artikelen fouten bevatten en wetenschappers vaak ongenueanceerde conclusies trekken wordt het lastig die informatie nog langer serieus te nemen. Men zou het immers ook moeilijk vinden de stabiliteit van de bouwsels van timmerlieden nog langer te vertrouwen wanneer bekend zou worden gemaakt dat ze hun gereedschap structureel verkeerd gebruikten. Dit is echter iets te kort door de bocht: je kunt je als lezer met enige moeite redelijk tegen deze interpretatievalkuilen wapenen. Hieronder worden daarom enkele suggesties beschreven die je als lezer kunnen behoeden in die valkuilen te stappen waar al zo velen in gestapt zijn.

### **Wantrouw claims van de afwezigheid van een effect**

Zoals niet te bewijzen is dat er geen oranje zwanen bestaan, zo is het bijna onmogelijk om op basis van een steekproef te

beweren dat er in de populatie geen effecten of verschillen zijn. Waar nog redelijk aannemelijk te maken is dat een effect niet afwezig is, is eigenlijk iedere steekproefgebaseerde uitspraak over de afwezigheid van een effect onzinnig. Hetzelfde geldt voor uitspraken als 'er is geen verschil tussen groep A en B', of 'A en B scoren hetzelfde op toets X'. Het feit dat je als auteur graag die uitspraak zou willen kunnen doen betekent helaas niet dat dit ook te verantwoorden is.

### **Bekijk zelf of de gevonden grootte van een effect interessant lijkt**

Stel dat iemand een significant verschil zou hebben gevonden tussen mannen en vrouwen voor wat betreft hun intelligentie: mannen blijken gemiddeld een IQ van 99.95 te hebben, en vrouwen een gemiddeld IQ van 100.05. Hoewel significant lijkt dit verschil van 0.10 IQ-punten in de praktijk in verreweg de meeste gevallen niet interessant of van belang. Als dit verschil echter 5 punten zou zijn zou het wel degelijk een belangrijke vondst zijn. Kortom: *significant is niet per se relevant!* Laat je als lezer daarom niet overtuigen door het woord significant, maar let ook op de grootte van een effect, en probeer dit op waarde te schatten. Kijk, indien gegeven, ook naar het betrouwbaarheidsinterval: dit geeft een redelijke indicatie waar het populatie-effect zich ongeveer bevindt. Een breed interval geeft hierbij aan dat je niet zo'n goede inschatting van die waarde kan maken, terwijl een relatief smal interval aangeeft dat deze waarde redelijk nauwkeurig te schatten lijkt.

### **Bewaar altijd een academische houding: neem nooit klakkeloos iets aan**

Hoe stellig een wetenschappelijk artikel ook geschreven is, uiteindelijk moeten het de uitkomsten zijn die overtuigen. Probeer daarom een artikel altijd met enige afstand te lezen, en ga er van uit dat de werkelijkheid niet zo zeker hoeft te zijn als de schrijver doet voorkomen. Een goede richtlijn is het volgende: 'extraordinary claims require extraordinary data'. Met andere woorden: laat je overtuigen door de gegevens, en niet door de stelligheid van de schrijver.

De bovenstaande suggesties vergen een kritische houding van de lezer, en dit is lang niet altijd eenvoudig. Vaak ben je als lezer minder expert dan de auteur van het artikel, en is de neiging dat

diegene het wel beter zal weten logisch en ook terecht. Aan de andere kant ligt de bewijslast volledig bij die auteur: hij of zij zal de lezer moeten proberen te overtuigen, en dat overtuigen zou moeten gebeuren door de gegevens en niet (alleen) door de manier waarop conclusies over die gegevens zijn geformuleerd.

Een probleem is dat je bij het lezen van een artikel vaak op zoek bent naar antwoorden op bepaalde vragen, en wanneer die antwoorden dan gegeven worden is het erg lastig deze vervolgens weer te relativiseren. De behoefte om een antwoord te krijgen wint het meestal van de behoefte om een juist of verantwoord antwoord te krijgen. De neiging om een korte en duidelijke uitspraak te doen is uitermate begrijpelijk: het is natuurlijk veel fijner om aan te kunnen geven hoe het al dan niet zit, dan om te zeggen dat de data in een bepaalde richting wijzen, maar dat niet helemaal zeker is of dit ook daadwerkelijk zo is. Toch zijn uiteindelijk zowel de wetenschap als ook de praktijk meer gebaat bij juistere maar voorzichtigere antwoorden op onderzoeksvragen, dan bij te stellige maar misschien onjuiste antwoorden. ■

#### **Literatuur**

- American Psychological Association (2001). *Publication Manual of the American Psychological Association* (5th ed.). Washington D.C.: Author.
- American Psychological Association (2009). *Publication Manual of the American Psychological Association* (6th ed.). Washington D.C.: Author.
- Berkson, J. (1938). Difficulties of interpretation encountered in the application of the Chi-square test. *Journal of the American Statistical Association*, 33, 526-536.
- Cohen, J. (1994). The earth is round ( $p < .05$ ). *American Psychologist*, 49, 997-1003.
- Hoekstra, R. (2009). *The use and usability of inferential techniques*. Ongepubliceerd proefschrift, Rijksuniversiteit Groningen, Groningen.
- Kline, R. B. (2004). *Beyond significance testing: Reforming data analysis methods in behavioral research*. Washington D.C.: American Psychological Association.
- Rosenthal, R. (1979). The file drawer problem and tolerance for null results. *Psychological Bulletin*, 86, 638-641.
- Thompson, B. (2007). Effect sizes, confidence intervals, and confidence intervals for effect sizes. *Psychology in the Schools*, 44, 423-432.
- Wilkinson, L., & APA Task Force on Statistical Inference. (1999). *Statistical methods in psychology journals: Guidelines and explanations*. *American Psychologist*, 54, 594-604.

■ Rink Hoekstra studeerde Psychologie en Technische Cognitiewetenschap in Groningen. Op 8 oktober 2009 promoveerde hij op zijn proefschrift over gebruik en bruikbaarheid van statistische technieken. Tegenwoordig is hij werkzaam als docent/onderzoeker bij het Gronings Instituut voor Onderzoek van Onderwijs, verbonden aan de afdeling Pedagogiek en Onderwijskunde aan dezelfde universiteit.