Replication Paradox

Met Michèle Nuijten

Michèle Nuijten studeerde psychologie aan de Universiteit van Amsterdam. In haar promotietraject aan het Methoden en Technieken van Onderzoek Departement aan de Universiteit van Tilburg richt zij zich op het detecteren en voorkomen van statistische fouten en datamanipulatie in psychologisch onderzoek. Zo heeft zij bijvoorbeeld het programma Statcheck ontwikkeld (te gebruiken in R), dat automatisch de juistheid van p-waardes in een artikel kan controleren. Daarnaast onderzoekt zij de "Replicatie Paradox", ofwel hoe slecht uitgevoerde replicatie studies juist meer kwaad dan goed kunnen doen.

Michèle Nuijten

"Het lijkt mij fantastisch als we middels empirische data inzicht kunnen krijgen in het gedrag van de wetenschapper en de invloed van wetenschappelijke fouten op de literatuur", beantwoord Michèle op de vraag waar haar PhD project over gaat. Deze interesse in de menselijke factoren in statistiek en goede wetenschapsbeoefening ontstond al snel, tijdens de bachelor Psychologie, waar ze zich al specialiseerde op methodologie. "Bij een vak over kritisch denken kregen we een aantal empirische artikelen voorgeschoteld die wemelden van de fouten, zowel theoretisch, als methodologisch en statistisch. Het was toen een beetje de sport om zoveel mogelijk fouten in een artikel aan te wijzen", grapt Michèle. "Maar al vrij snel realiseerde ik me dat je daar niet echt mee verder komt".

Het is heel gemakkelijk om in gepubliceerde artikelen fouten te vinden, want het doen van onderzoek is complex. Dat er per ongeluk fouten gemaakt kunnen worden is te begrijpen, maar wat daarin wel heel belangrijk is, is dat de fouten die we maken willekeurig zijn en niet resulteren in systematische *bias*. Helaas is dat nu niet het geval en dus moet daar verandering in komen. Maar wat kan hier precies aan gedaan worden? En hoe erg zijn de problemen eigenlijk? "Ondanks dat er veel mensen zijn die hier ideeën over hebben, is er ironisch genoeg verrassend weinig empirisch onderzoek gedaan naar wetenschap zelf. Juist dat is waar ik mij nu in mijn promotietraject mee bezig houd".

Een belangrijk onderdeel van het project is het onderzoek naar de "replicatie paradox". Stel je voor; Je bent bezig met het opzetten van een nieuwe studie, waarin je een bepaald fenomeen wilt onderzoeken. Je schrijft een ethische toestemmingsaanvraag. Een belangrijk aspect daarvan is de "*power* analyse". Hoeveel deelnemers heb je nodig om de door jou verwachte *effect size* te kunnen bereiken? En hoe moet je in de eerste plaats een inschatting maken van deze ongrijpbare effect size?

Elke wetenschapper zal deze situatie herkennen. Voor velen zal deze analyse ook worden ervaren als een enigszins irritant, maar verplicht onderdeel. De procedure is standaard: je duikt in de literatuur, op zoek naar vergelijkbare onderzoeken die het fenomeen hebben onderzocht en maakt een schatting van de in de literatuur gevonden effect sizes. Vervolgens zijn er programmaatjes waar je deze gegevens invult en voilà: ik weet hoe groot mijn steekproef zou moeten zijn en zet dit in mijn aanvraag.

"Precies, dat is hoe het overgrote deel van de wetenschappers het probleem van de effect sizes en steekproefgroottes aanpakt" beaamt Michèle. "Dat is op zich niet verkeerd, maar uit ons onderzoek blijkt dat wij misschien niet zo heel goed in staat zijn om de juiste effect size in te schatten! Laten we er een voorbeeld bij pakken. In je zoektocht naar literatuur heb je geluk: je vindt zowel een grote studie als een

kleine replicatiestudie. Hoe interpreteer je deze gegevens om een schatting te maken van de algemene effect size?"

Michèle besloot deze vraag voor te leggen aan een grote groep studenten én wetenschappers uit zowel toegepaste sociale wetenschappen als psychometrie en methodologie. De deelnemers beoordeelden welke combinaties van grote en kleine studies de beste schatting van het effect van een behandeling in een algemene populatie zou opleveren. De resultaten waren niet verrassend. De meerderheid van de studenten en wetenschappers, onafhankelijk van hun expertise, kiezen ervoor om beide studies te gebruiken om een algemene effect size te schatten. Wetenschappers kozen hierin als groep trouwens niet anders dan de studenten.

Intuïtief geven we dus de voorkeur aan het gebruiken van beide studies. Deze redenering is logisch, want hoe meer informatie, hoe beter. Toch? "Nee dus", zegt Michèle. "Helaas heeft onze intuïtie het hier fout!".

Hoezo dan? Er wordt tegenwoordig een grote nadruk gelegd op het belang van replicatiestudies. Gebaseerd op een grote hoeveelheid literatuur, pleit men voor meer replicatiestudies, omdat die de schattingen nauwkeuriger zouden maken en het aantal *false positives* in onderzoek zouden verminderen. Helaas zijn er twee fenomenen die deze kwestie ingewikkelder maken dan het op het eerste gezicht lijkt, namelijk publicatie bias en power.

Elke wetenschapper kent het begrip 'publicatiebias'. Studies met een significant resultaat hebben een veel grotere kans om gepubliceerd te worden. De nadelen van een dergelijke trend zijn duidelijk, want effect size wordt opgeblazen als alleen significante effecten in tijdschriften worden geaccepteerd. Dit is een zorgwekkend maar alom bekend fenomeen, waar ook veel aandacht aan wordt besteed.

"Ik hoor je denken: daarom doen we toch replicatie studies? Dat klopt! Helaas is er bewijs uit meta-analyses en artikelen die meerdere studies bevatten dat óók replicatiestudies onderhevig zijn aan publicatiebias", verklaart Michèle. Dat betekent dus dat zowel gepubliceerde studies als gepubliceerde replicaties overschatte effect sizes bevatten! Het resultaat hiervan is dat als je de resultaten van een gepubliceerde originele studie combineert met die van een gepubliceerde replicatie, de schatting dus nog minder nauwkeurig wordt dan wanneer je die baseert op maar één studie.

Dit werkt als volgt. De bias in de effect size schatting in één studie hangt af van zowel de mate van publicatie bias als de power. Studies met een lage power zullen een minder betrouwbare effect size schatting tot gevolg hebben, die zowel tot een zware onderschatting als een overschatting van de ware effect size kan leiden. Gelukkig leidt het bepalen van een gemiddelde van al deze *underpowered* studies tot een redelijk nauwkeurige schatting van de werkelijke effect size. "Maar dat is alleen het geval als er géén publicatie bias is. Is die er wel, dan eindigen alleen de sterke overschattingen in de literatuur, want die zullen significant zijn", voegt Michèle toe. "Het combineren van alle gepubliceerde, significante effect sizes leidt dan dus tot een zwaar opgeblazen gemiddelde".

Aan de andere kant hebben studies met een hoge power dit probleem niet. De effect sizes uit deze studies zijn preciezer en liggen dichter bij de werkelijke effect size. Dat houdt dus in dat de gemiddeld geschatte effect size dus niet zo wordt vervormd als deze wordt gebaseerd op deze studies, ondanks de publicatie bias.

Laten we met deze kennis in ons achterhoofd teruggaan naar het scenario dat we hebben voorgelegd aan al die wetenschappers, waar een grote studie en een kleine replicatie studie beoordeeld moeten worden. Hoe evalueer je deze informatie? Michèle geeft het juiste antwoord: "Als we er van uitgaan dat beide studies onderhevig zijn geweest aan publicatie bias, dan is de effect size van de grote studie dus waarschijnlijk lichtelijk te hoog. De effect size van de kleine replicatie studie is daarentegen waarschijnlijk veel zwaarder overschat. In dat geval zal je schatting dus nauwkeuriger zijn als je de replicatie studie niet meeneemt in je oordeel!"

Kortom: een replicatiestudie zal de **precisie** vergroten (het betrouwbaarheidsinterval rond de effect size zal kleiner worden), maar replicatie zal de **bias** vergroten als de sample size kleiner is dan de originele studie, **mits** er sprake is van publicatie bias en lage power.

De meest voor de hand liggende oplossing voor dit probleem is zorgen dat we afkomen van die publicatie bias. Het probleem is dan meteen opgelost: als er geen publicatie bias is, dan is er geen systematische overschatting van de effecten. Gelukkig zijn er steeds meer initiatieven voor replicatiestudies op grote schaal die op voorhand worden vastgelegd en worden gepubliceerd ongeacht de uitkomst. Ook zijn er tijdschriften zoals PLOS ONE die expliciet stellen alleen de methodologische kwaliteit van een onderzoek te beoordelen, en de beslissing om een stuk te publiceren niet af laten hangen van de resultaten.

Maar wat te doen met al die studies die al gepubliceerd zijn? “Om daar mee om te gaan is het belangrijk dat je alleen studies meeneemt in je schatting die een hoge power hebben. En dat je studies met lage power dus negeert” vat Michèle samen. Aangezien we niet precies weten hoe groot de rol van publicatie bias is in de wetenschappelijke literatuur, maar het er wel op lijkt dat het alom aanwezig is, is het misschien wijs als wetenschappers uitgaan van een *worst-case scenario*. De houding van “hoe meer informatie, hoe beter” moet worden veranderd naar “**Hoe meer power, hoe beter**”.

Dat is een heldere boodschap, maar wellicht ook makkelijker gezegd dan gedaan. Wat doe je als je maar geld hebt voor een kleine studie? “Dat is inderdaad de vraag die ik steevast krijg”, zegt Michèle, “en het is voor mij ook een stuk makkelijker praten, aangezien ik als methodoloog in mijn onderzoek weinig te maken heb met proefpersonen. Toch is het van essentieel belang dat studies een hoge power hebben. De problemen die studies met een lage power opleveren zijn niet alleen theoretisch methodologengezeur, maar hebben vergaande gevolgen voor de interpretatie van onderzoek, wat op de lange termijn een effect kan hebben op bijvoorbeeld maatschappelijke toepassingen of bijvoorbeeld beleid”. Een mogelijke oplossing voor onderzoekers die maar budget hebben voor een kleine studie is om de krachten te bundelen met andere onderzoeksgroepen. Een bijkomend voordeel hiervan is dat de onderzoekers dan waarschijnlijk beter hun onderzoeksmethoden en data zullen documenteren – een andere onderzoeksgroep moet het immers ook kunnen begrijpen – waardoor er minder snel fouten in het onderzoek zullen sluipen en het makkelijker wordt het onderzoek eventueel te repliceren.

Het probleem is alleen dat het huidige wetenschappelijke systeem dit soort samenwerkingsgedrag niet beloont. Integendeel, in het huidige systeem word je niet beloond voor een nauwkeurige effect size schatting, maar voor significante resultaten. En hoewel het misschien tegen-intuïtief klinkt, heb je de grootste kans op significante resultaten als je veel kleine samples draait in plaats van één grote. In een kleine sample heb je meer ruis in je meting, en een grotere kans dat je per ongeluk een significant effect vindt. Het gevolg daarvan is dat wetenschappers strategisch gedrag gaan vertonen, en het geld voor één grote sample opsplitsen om kleine samples van te draaien, om zo hun publicatiekansen te verhogen. Een samenwerkingsverband aangaan met een andere groep is op die manier dus helemaal niet in het persoonlijke belang van de wetenschapper.

Als tijdschriften inderdaad alleen maar significante resultaten willen, is het voor wetenschappers niet strategisch om grote samples met hoge power te draaien. “Maar wat blijkt nu: de grootste selectie van significante resultaten vindt niet plaats bij de tijdschriften, maar bij de auteurs zelf”, stelt Michèle. “Veel wetenschappers die een niet significant resultaat vinden, nemen niet meer de moeite om het artikel te schrijven en op te sturen!” Ook door dit fenomeen verschijnen er meer significante resultaten in de literatuur, wat weer leidt tot overschattingen van effecten. Het zou dus al helpen als wetenschappers het artikel schrijven, ongeacht de uitkomst van de studie.

Verder schreeuwt het huidige wetenschappelijke systeem om een ander beloningsmechanisme: je zou niet moeten

worden beloond voor het aantal (significante) studies dat je hebt gepubliceerd, maar voor de wetenschappelijke kwaliteit van je werk. Dat klinkt behoorlijk abstract, maar er zijn al wat pogingen hiertoe. Zo kan je bijvoorbeeld bij de tijdschriften van The Association for Psychological Science badges krijgen voor het delen van je data, het delen van je materiaal, en het pre-registreren van je studie. Deze badges komen bij je artikel te staan, en zijn als het ware een kwaliteitskeurmerk. Het zal nog wel even duren voordat het hele systeem dit soort beloningen uitdeelt, maar het biedt goede vooruitzichten.

"Het moeilijkste aan dit vakgebied is dat we fouten van wetenschappers onderzoeken, terwijl we zelf ook wetenschappers zijn met dezelfde menselijke fouten en vooroordelen", geeft Michèle toe. "We zijn dan ook constant bezig met manieren verzinnen om te zorgen dat we niet in dezelfde valkuilen trappen, wat voor mij het werk heel erg concreet en interessant maakt. Naast het empirisch onderzoeken van deze vraagstukken, in de hoop om op de lange termijn een zinvolle bijdrage te kunnen leveren aan bijvoorbeeld het beleid van tijdschriften, houden Michèle en haar collega's van de onderzoeksgroep zich met liefde bezig met wetenschappers bewuster maken van deze kwesties. "Vaak krijg ik op congressen verbaasde reacties op dit verhaal. Ik ben blij dat ik ook op deze kleine schaal mijn steentje bij kan dragen, door dit probleem op een begrijpelijke wijze uit te leggen. Maar mijn grote hoop is dat dit soort onderzoek daadwerkelijk een verandering teweeg gaat brengen in hoe de wetenschap op dit moment werkt en afgaande op de initiatieven die er nu al zijn, zijn we op de goede weg."

Dit stuk is gebaseerd op het artikel: "The Replication Paradox: Combining Studies Can Decrease Accuracy of Effect Size Estimates", door Michèle B. Nuijten, Marcel A. L. M. van Assen, Coosje L. S. Veldkamp, and Jelte M. Wicherts, Tilburg University (submitted).

Zowel de wetenschappelijke poster over dit project, gepresenteerd tijdens het seminar "Improving scientific practice, dealing with the human factors" te Amsterdam, als meer informatie over Statcheck is te vinden op http://mbnuijten.com.