REPLICEERBAARHEID IN DE EMPIRISCHE MENSWETENSCHAPPEN

Patrick Onghena

KVAB STANDPUNTEN
64
Koninklijke Vlaamse Academie van België
voor Wetenschappen en Kunsten - 2020

REPLICEERBAARHEID IN DE EMPIRISCHE MENSWETENSCHAPPEN

KVAB Press

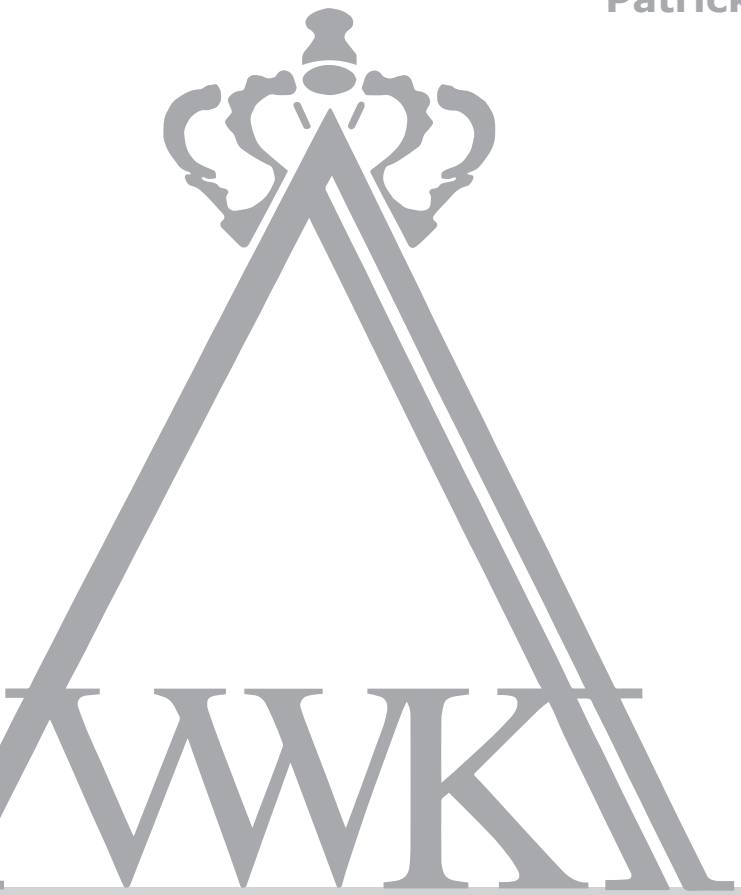
KVAB STANDPUNTEN
64

Concept cover: Francis Strauven
Ontwerp cover: Charlotte Dua
Afbeelding: Shutterstock

De tekening van het Paleis der Academiën is een reproductie van het originele perspectief van Charles Vander Straeten in 1823. Jozef Cantré ontwierp het logo van de KVAB in 1947.
De KVAB Standpunten worden gepubliceerd door de Koninklijke Vlaamse Academie van België voor Wetenschappen en Kunsten, Hertogsstraat 1, 1000 Brussel.
Tel. 00 32 2 550 23 23 – info@kvab.be – www.kvab.be

REPLICEERBAARHEID IN DE EMPIRISCHE MENSWETENSCHAPPEN

Patrick Onghena



D/2020/0455/01
ISBN 978 90 656 919 41

Aanbevolen citeerwijze: Patrick Onghena, *Repliceerbaarheid in de empirische menswetenschappen*, KVAB Standpunt 64, 2020.

Drukkerij Universa

Repliceerbaarheid in de empirische menswetenschappen

INHOUD

Samenvatting

Van wetenschappelijke bevindingen verwachten we dat ze repliceerbaar zijn. Deze verwachting houdt in dat onafhankelijke onderzoekers dergelijke bevindingen in onafhankelijk vervolgonderzoek kunnen herhalen. Is deze verwachting realistisch voor wetenschappelijk onderzoek in de empirische menswetenschappen? Niet altijd, zo blijkt uit een recente, indrukwekkende reeks vervolgonderzoeken. Wetenschappelijke bevindingen in de empirische menswetenschappen blijken in onafhankelijk vervolgonderzoek niet altijd stand te houden. Als er toch een effect wordt gevonden, blijkt dat bovendien meestal kleiner te zijn dan in de oorspronkelijke studie. Zitten we, behalve met *fake news*, nu ook met *fake science* opgescheept?

In dit Standpunt gaan we in op de uitgangspunten en het begrippenkader van deze zogenaamde 'repliceerbaarheidscrisis'. We onderzoeken de antecedenten en mogelijke oorzaken van de beperkte repliceerbaarheid en bespreken de implicaties voor de geloofwaardigheid van de empirische menswetenschappen. Op basis van een analyse van de oorzaken en de implicaties trachten we een weg uit de crisis uit te stippelen. Zoals psychologen beweren: elke crisis bevat de kiem voor groei. We willen graag hopen dat deze groei bij de huidige crisis repliceerbaar is.

Executive summary

Replicability in the empirical social, behavioral, and educational sciences

We expect scientific findings to be replicable, i.e. that they can be repeated by independent researchers in independent follow-up studies. Is this standard feasible for academic research in the empirical social, behavioral, and educational sciences? Not always, according to a recent impressive series of follow-up studies. Findings in these sciences don't always appear to hold up in independent follow-up studies. Moreover, if an effect is found in the follow-up study, it is usually smaller than in the original study. After *fake news*, do we now have to put up with *fake science*?

In this position paper we take a look at the basic principles and the conceptual framework behind this so-called 'replicability crisis'. We investigate the antecedents and the possible causes of the limited replicability and discuss the implications for the credibility of the empirical social, behavioral, and educational sciences. On the basis of an analysis of the causes and implications, we endeavour to chart a course through this crisis. As the psychologists say: in every crisis lies the seed of growth. We would like to think that this growth is replicable in the current crisis.

Voorwoord

Reeks Standpunten

De reeks Standpunten van de Academie is een bijdrage tot een wetenschappelijk onderbouwd debat over actuele maatschappelijke en artistieke thema's. De auteurs, leden en werkgroepen van de Academie schrijven in eigen naam, onafhankelijk en met volledige intellectuele vrijheid. De goedkeuring voor publicatie door een of meerdere Klassen van de Academie waarborgt de kwaliteit van de publicatie. Dit Standpunt werd goedgekeurd voor publicatie door de Klasse van de Menswetenschappen op 18 januari 2020.

Verantwoording

Dit Standpunt is gebaseerd op de mondelinge mededeling *De repliceerbaarheidscrisis in de empirische menswetenschappen*. Die werd op 17 november 2018 gebracht voor de Klasse van de Menswetenschappen van de Koninklijke Vlaamse Academie van België voor Wetenschappen en Kunsten. Ze is aangevuld met informatie uit de meest relevante publicaties van 2019, waaronder het rapport over *Reproducibility and Replicability in Science* van de Amerikaanse National Academies of Sciences, Engineering, and Medicine (2019).

Dankwoord

Graag dank ik de leden van de Klasse voor de boeiende en kritische nabespreking, net als de vast secretaris van de Academie, Freddy Dumortier, en de voorzitter en de ondervoorzitter van de Klasse, Kristiaan Versluys en Jo Tollebeek, voor het vertrouwen en de aanmoedigingen om de mededeling tot een Standpunt uit te werken. Confrater Johan Wagemans dank ik voor het grondige naleeswerk en de constructieve suggesties bij de eindversie van de tekst.

Bij de totstandkoming van dit Standpunt hebben de leden van de Jonge Academie een vooraanstaande rol gespeeld. Ik wil hen van harte danken voor de richtinggevende ideeën en adviezen in de loop van het schrijfproces, en voor de uitgebreide feedback achteraf. In het bijzonder gaat mijn dank naar Sylvia Wenmackers, Vincent Ginis, Christophe Vandeviver, Jozefien De Leersnyder en Bert De Smedt.

Mijn grootste dank gaat naar mijn vrouw, Charlotte Struyve, die zowel mijn intellectueel klankbord is als mijn grootste inspiratiebron én mijn steun om alles aan het thuisfront te beredderen terwijl ik aan dit Standpunt werkte. Ik draag dit Standpunt graag op aan onze kinderen, Victor en Alice, twee geslaagde systematische replicaties.

1. Probleemstelling

Er waait een repliceerbaarheidsstorm door de empirische menswetenschappen. Bevindingen die tot de canon van vele disciplines werden gerekend en ondertussen als kernleerstof in algemene handboeken zijn opgenomen, blijken moeilijker repliceerbaar dan aanvankelijk gedacht (Ferguson, Brown, & Torres, 2018; Lilienfeld & Waldman, 2017; Pashler & Harris, 2012). Als er in de replicatiestudies een effect wordt gevonden, blijkt dat bovendien meestal kleiner te zijn dan in de initiële studie (Camerer et al., 2016, 2018; Ebersole et al., 2016; Klein, 2014, 2018, 2019; Open Science Collaboration, 2015).

In dit Standpunt gaan we in op de uitgangspunten en het begrippenkader van deze 'repliceerbaarheidscrisis'. We onderzoeken de antecedenten en mogelijke oorzaken van de beperkte repliceerbaarheid en bespreken de implicaties voor de geloofwaardigheid van de empirische menswetenschappen. Op basis van een analyse van de oorzaken en de implicaties trachten we een weg uit de crisis uit te stippelen. Zoals psychologen beweren: elke crisis bevat de kiem voor groei. We hopen dat deze groei bij de huidige crisis repliceerbaar is.

In zekere zin is dit Standpunt een vervolg op het KVAB-Standpunt 62, *De strijd om de waarheid. Over nepnieuws en desinformatie in de digitale mediawereld* (Billiet, Opgenhaffen, Pattyn, & Van Aelst, 2018). In het verlengde van de vragen 'kunnen we de media nog vertrouwen als het over feiten en informatie gaat?', en 'hoe verdedigen de traditionele media zich tegenover nepnieuws?' liggen deze vragen: 'kunnen we de wetenschap nog vertrouwen als het over feiten en informatie gaat?', en 'hoe verdedigt de wetenschap zich tegenover nepwetenschap?' Of om het radicaler uit te drukken: 'Wat is een feit?'.[1] Ook binnen de wetenschap woedt de strijd om de waarheid.

Vermeldenswaardig is trouwens dat in het Standpunt over nepnieuws wordt verwezen naar repliceerbaarheid, naar aanleiding van de bespreking van de jaarrede van José van Dijck, toen zij in 2018 afscheid nam als voorzitster van de Koninklijke Nederlandse Akademie van Wetenschappen. Over het vertrouwen in de wetenschap (in vergelijking met de media) wordt daar het volgende gesteld:

> 'Over een aantal feiten is na jarenlang grondig onderzoek, kritische reflectie en discussie binnen de wetenschappelijke disciplines een grote consensus ontstaan. Voor veel onderwerpen is dat echter (nog) niet het geval. Volgens Van Dijck hoeven die onzekerheid en twijfel het vertrouwen in de wetenschap niet te ondermijnen, zolang het proces van wetenschappelijke kennisverwerving gericht is op het vinden van een *common ground*: een

[1] Over deze vraag bestaat overigens interessant recent historisch onderzoek. Zie bijvoorbeeld Ten Hagen (2019).

> verzameling feiten en inzichten die op een zorgvuldige wijze tot stand is gekomen. "Zorgvuldigheid" verwijst hier naar een proces dat beantwoordt aan de eisen van integriteit, transparantie, onafhankelijkheid en rekenschap afleggen. Onderzoekers beantwoorden aan die vereisten door met elkaar samen te werken en te communiceren, en vooral ook door elkaars maat te nemen en tegenspraak te organiseren, onder meer via replicatieonderzoek.' (Billiet et al., 2018, p. 23)

Uit dit citaat blijkt de prominente positie die repliceerbaarheid inneemt als het over vertrouwen in de wetenschap en zorgvuldigheid in wetenschappelijk onderzoek gaat. Het citaat houdt ook een waarschuwing in. Nieuwsberichten zijn noodzakelijkerwijs snel en gericht op 'het nieuwe'. Als we ons in het wetenschappelijk onderzoek met dezelfde snelheid zouden richten op sensatie en originaliteit (*scoops*) – ten koste van consensusvorming, zorgvuldigheid en consolidering – dan bestaat het risico dat we, behalve met *fake news*, ook met *fake science* opgescheept worden (zie bv. Brainard, 2018; Bucci, 2019; Fang & Casadevall, 2011; Frith, 2020).

2. Plaatsbepaling en begrippenkader

2.1. Plaatsbepaling

Dit Standpunt onderzoekt de repliceerbaarheid in de *empirische menswetenschappen*. De beperking tot *mens*wetenschappen is ingegeven door overwegingen van haalbaarheid en focus. Omdat het Standpunt geformuleerd werd binnen de Klasse van de Menswetenschappen van de Koninklijke Vlaamse Academie van België voor Wetenschappen en Kunsten, ligt deze beperking voor de hand. Het Standpunt houdt evenwel ook een uitnodiging in om vervolgstandpunten in te nemen voor de medische, de technische en de natuurwetenschappen. Zoals uit de uiteenzetting zal blijken, zijn er veel parallelle inzichten over repliceerbaarheid tussen de verschillende wetenschapsdomeinen mogelijk. In een bepaald opzicht kan zelfs worden gesteld dat de 'repliceerbaarheidscrisis' binnen de *medische* wetenschappen tot de eerste volledige uitbraak gekomen is (zie onder 3 over de antecedenten, stand van zaken en mogelijke oorzaken).

De beperking tot *empirische* wetenschappen heeft te maken met het onderwerp van dit Standpunt enerzijds en de breedte van disciplines die onder de menswetenschappen kunnen worden gerekend anderzijds. Het Standpunt gaat over repliceerbaarheid en dit impliceert empirische operationalisering, observeerbaarheid en herhaalbaarheid (zie 2.2. over de definitie). Niet voor alle menswetenschappen is repliceerbaarheid even relevant – denk aan de theologie, de filosofie, de rechtswetenschap, de literatuurwetenschap, het historisch onderzoek, bepaalde delen van de politicologie, de pedagogische wetenschappen en de sociale en culturele antropologie – of zou het een herconceptualisering van repliceerbaarheid vragen die de verdere bespreking zou compliceren.

Zelfs binnen de empirische menswetenschappen kan er nog een verdere beperking worden aangebracht: het repliceerbaarheidsvraagstuk rijst voornamelijk bij hypothesetoetsend onderzoek in de gedrags- en maatschappijwetenschappen (De Groot, 1961; Wagenmakers, Wetzels, Borsboom, van der Maas, & Kievit, 2012). Onderzoek dat gericht is op een eerste verkenning of op ontdekking (exploratief onderzoek en *discovery science*; zie bv. Biswal, 2010) en studies die cultuurwetenschappelijke methoden hanteren (hermeneutiek en conceptueel onderzoek; zie bv. Ravitch & Carl, 2016), vallen buiten het bestek van dit Standpunt, waarmee we ons dus eerder in een *context of justification* dan in een *context of discovery* situeren (Hoyningen-Huene, 1987; Reichenbach, 1938).

Deze beperkingen lijken tot een Standpunt met een smal bereik te leiden. In verhouding tot de volledige breedte van de menswetenschappen is dat zeker het geval, maar in absolute termen gaat het nog over een gigantisch domein, gaande van de meest dominante benaderingen in de psychologie, over belangrijke onderdelen van de pedagogische wetenschappen, de sociale wetenschappen en

de criminologie, tot het experimenteel economisch onderzoek. Repliceerbaarheid is overigens niet alleen relevant voor (fundamenteel en toegepast) experimenteel onderzoek, maar ook voor observationeel onderzoek (bv. het onderzoek naar de taalontwikkeling bij kinderen, Bergmann et al., 2018) en voor correlationeel onderzoek (bv. het onderzoek naar het verband tussen persoonlijkheid en belangrijke levensuitkomsten, zoals huwelijksstabiliteit, beroepsgericht succes, politieke overtuigingen en strafregister, Soto, 2019). Taalontwikkeling is bijvoorbeeld operationaliseerbaar en in die zin herhaaldelijk observeerbaar bij elk opgroeiend kind. Hetzelfde geldt voor persoonlijkheid en belangrijke levensuitkomsten: persoonlijkheid wordt niet gemanipuleerd in dit type van onderzoek, maar er kan wel herhaaldelijk, en in een verscheidenheid aan steekproeven, nagegaan worden of er een verband is tussen de persoonlijkheid van de onderzoeksdeelnemers en latere levensuitkomsten en, als er een verband wordt gevonden, hoe groot dat verband is.

Met deze laatste twee voorbeelden van onderzoek in de empirische menswetenschappen willen we nog aangeven dat een groot deel van dit Standpunt weliswaar is geformuleerd in algemeen methodologische termen, maar dat de relevantie van repliceerbaarheid moet worden beoordeeld binnen specifieke sets van wetenschappelijke toepassingen. Het gaat met name over onderzoek waarbij de onderzoekers er zelf impliciet of expliciet van uitgaan dat hun bevindingen een mate van algemene geldigheid bezitten en dus herhaalbaar moeten zijn. Om de relevantie van repliceerbaarheid voor specifieke sets van wetenschappelijke toepassingen te illustreren, gaan we in dit Standpunt dieper in op twee casussen: het voedingspsychologisch onderzoek van Brian Wansink in 3.3.3 en het onderzoek over lichaamstaal van Amy Cuddy in 4.1.

2.2. Begrippenkader

Repliceerbaarheidsdiscussies leiden soms tot misverstanden als termen door elkaar worden gehaald (National Academies of Sciences, Engineering, & Medicine, 2019; Plesser, 2018). In een eerste paragraaf (2.2.1.) stellen we daarom een definitie van de termen 'replicatiestudie', 'repliceerbaarheid' en 'replicatie' voor. In een tweede paragraaf (2.2.2.) bakenen we een aantal aanverwante begrippen af.

2.2.1. Replicatiestudie, replicatie en repliceerbaarheid

Peels (2019) maakt een onderscheid tussen een 'replicatiestudie', 'replicatie' en 'repliceerbaarheid'. We volgen hem hierin en combineren het met de omschrijvingen die ook in het recente Amerikaanse rapport over *Reproducibility and Replicability in Science* van de National Academies of Sciences, Engineering, and Medicine (2019) terug te vinden zijn:

Replicatiestudie

Een replicatiestudie is een onafhankelijke herhaling van een eerdere studie, gebaseerd op nieuw verzamelde empirische gegevens en gebruik makend van zo veel mogelijk gelijkaardige methoden in zo gelijkaardig mogelijke omstandigheden als in de eerdere studie (National Academies of Sciences, Engineering, & Medicine, 2019; Peels, 2019).

Replicatie

Een replicatiestudie kan als een replicatie worden beschouwd als de resultaten ervan consistent zijn met de resultaten van de eerdere studie, rekening houdend met de onzekerheid die eigen is aan empirisch onderzoek (National Academies of Sciences, Engineering, & Medicine, 2019; Peels, 2019).

Repliceerbaarheid

Repliceerbaarheid verwijst naar een verzameling van eigenschappen van een studie die het mogelijk maken om een replicatiestudie uit te voeren (Peels, 2019). *Repliceerbaarheid* wordt ook gebruikt als algemene term om te verwijzen naar de mogelijkheid en/of aanwezigheid van *replicaties* over een bepaald fenomeen of in een bepaald onderzoeksdomein (National Academies of Sciences, Engineering, & Medicine, 2019).

Bij deze definities kunnen we een aantal kanttekeningen maken.

Kanttekeningen bij de definitie van 'replicatiestudie'

In de definitie van 'replicatiestudie' zijn een aantal begrippen opgenomen die een subjectieve beoordeling vergen: 'onafhankelijke', 'zo veel mogelijk' en 'zo gelijkaardig mogelijk', 'gelijkaardige methoden' en 'omstandigheden'.

In navolging van Peels (2019) beschouwen we de replicatiestudie als 'onafhankelijk' als ze op geen enkele manier afhangt van de *resultaten* van de eerdere studie. Bij de proefopzet en argumentatie van de replicatiestudie mag dus niet worden verondersteld dat de resultaten van de eerdere studie betrouwbaar en geldig zijn. De gegevens mogen bovendien op geen enkele manier overlappen met de gegevens van de eerdere studie. Sommige aspecten van afhankelijkheid zijn vanzelfsprekend wel toegelaten, of zelfs noodzakelijk, om van een replicatiestudie te spreken. De replicatiestudie kan bijvoorbeeld wel gebruik maken van dezelfde onderzoeksinstrumenten en hetzelfde onderzoeksprotocol. In bepaalde gevallen kan het zelfs aangewezen zijn om de oorspronkelijke onderzoekers bij de replicatiestudie te betrekken of te consulteren.

De verwijzing naar 'zo veel mogelijk', 'zo gelijkaardig mogelijk', 'gelijkaardige methoden' en 'omstandigheden' in de definitie van 'replicatiestudie' geeft aan dat er variatie mogelijk is in de gelijkenis tussen de eerdere studie en de replicatiestudie. Bij replicatiestudies gaat het enerzijds telkens over dezelfde onderzoeksvraag (hypothesen en effecten) en zullen er anderzijds altijd verschillen in contextuele variabelen zijn, waarop de onderzoekers weinig of geen vat hebben: moment van afname en historische context, geografische locatie, cultuur, of taal. Voor andere methodologische dimensies kunnen er wél expliciete keuzes worden gemaakt. LeBel, McCarthy, Earp, Elson en Vanpaemel (2018) hanteren hiervoor een taxonomie die de graad van gelijkenis tussen de eerdere studie en de replicatiestudie aangeeft, variërend over die methodologische dimensies: de directe onderzoekssetting, de procedurele details, de gehanteerde stimuli voor de afhankelijke of de onafhankelijke variabelen, de onderzoekspopulatie (bv. leeftijd en geslacht), de operationaliseringen voor de onafhankelijke of afhankelijke variabelen, en de constructen voor de onafhankelijke of afhankelijke variabelen. Zij pleiten ervoor om alleen de studies die variaties in onderzoekssetting, procedure en stimuli aanbrengen (de zogenaamde *directe replicaties*) als bewijsvoering voor replicatie op te nemen. Bij de studies die variëren op vlak van onderzoekspopulatie, operationaliseringen en constructen (de zogenaamde *conceptuele replicaties*) ligt het namelijk te zeer voor de hand om een niet-replicatie toe te schrijven aan de ingrijpende methodologische verschillen.

Welk standpunt men over het belang van directe en conceptuele replicaties ook inneemt, het is duidelijk dat er een subjectieve beoordeling nodig is om een studie als een replicatiestudie te erkennen. Dit hoeft geen nadeel te zijn of een louter individuele beoordeling in te houden, maar maakt integraal deel uit van het wetenschappelijk debat (De Groot, 1961).[2] Uiteindelijk is een strikte demarcatie tussen een studie die het etiket 'replicatiestudie' verdient en een studie die dat niet verdient zelfs niet nodig. Repliceerbaarheidsdiscussies kunnen rekening houden met *de mate* waarin een studie als een replicatiestudie kan worden beschouwd, of met de *kwalitatieve* gelijkenissen en verschillen tussen de eerdere studie en de replicatiestudie.

Onderzoekers zouden wel moeten kunnen aangeven wat de kernkenmerken van de studie en van het onderzochte fenomeen zijn, en wat de randkenmerken. Randkenmerken zijn varieerbaar over replicatiestudies heen, kernkenmerken niet. Dit onderscheid is nodig omdat anders een replicatiestudie principieel onmogelijk wordt. De contextuele variabelen van tijd en ruimte die eerder werden vermeld,

[2] De Groot (1961) verwijst in dit verband naar het 'wetenschappelijk forum', maar het zou ons te ver leiden om verder in te gaan op de noodzakelijke intersubjectiviteit van wetenschap en op zijn forumtheorie. Het volstaat hier op te merken dat communicatie, tegenspraak en consensusvorming integraal deel uitmaken van wetenschappelijk onderzoek, zoals ook blijkt uit het citaat op bladzijde 10: 'Onderzoekers beantwoorden aan die vereisten door met elkaar samen te werken en te communiceren, en vooral ook door elkaars maat te nemen en tegenspraak te organiseren'.

zorgen er namelijk altijd voor dat onderzoekers 'nooit twee keer door dezelfde rivier kunnen waden' (Barsalou, 2016; Morawski, 2019; Van Bavel, Mende-Siedlecki, Brady, & Reinero, 2016). Ofwel zijn de contextuele variabelen relevant en moeten ze als kernkenmerken (of *moderatoren*) in het verklaringsmodel worden opgenomen, ofwel zijn het randkenmerken die er verder weinig of niet toe doen. Dit geldt trouwens ook voor detailkenmerken van de onderzoekssetting: van de verlichting en verwarming in het onderzoekslokaal tot 'de lengte van de baard van de onderzoeksleider' (Coyne, 2016).

Kanttekeningen bij de definitie van 'replicatie'

Vooreerst moet worden opgemerkt dat in onze definitie 'replicatie' als synoniem voor 'geslaagde replicatie' wordt gebruikt. Een andere manier om dat uit te drukken is zeggen dat in de replicatiestudie de resultaten van de eerdere studie zijn *gerepliceerd*. Bij een replicatiestudie die leidt tot resultaten die *inconsistent* zijn met de resultaten van de eerdere studie, rekening houdend met de onzekerheid die eigen is aan empirisch onderzoek, kan dan over 'niet-replicatie' worden gesproken.

Twee andere begrippen uit de definitie die toelichting vergen zijn 'consistentie' en 'onzekerheid'. Om te kunnen besluiten tot 'replicatie' moet er een vergelijking plaatsvinden tussen de resultaten van de eerdere studie en van de replicatiestudie. Die zullen nooit identiek zijn vanwege inherente onnauwkeurigheden in meet-instrumenten, de nieuwe gegevensverzameling met andere deelnemers en eventuele variaties in setting, procedure en stimuli. Er zijn echter meerdere manieren om de consistentie – of omgekeerd: het verschil tussen de resultaten van de eerdere studie en van de replicatiestudie – te operationaliseren. Bovendien is er dan nog altijd een afspraak nodig om een bepaald verschil als verwaarloosbaar te beschouwen (*consistentie*) en een ander als betekenisvol (*inconsistentie*). De 'onzekerheid' waarmee men rekening moet houden bij afwegingen van consistentie en die het resultaat is van onbetrouwbaarheid in zowel de eerdere studie als de replicatiestudie, steekproevenvariabiliteit en onderzoeksvariatie, krijgt ook een centrale plaats in het recente Amerikaanse rapport (National Academies of Sciences, Engineering, & Medicine, 2019). Replicatieonderzoek zal dus, net zoals oorspronkelijk onderzoek, zorgvuldig de operationalisering van consistentie, de criteria en de statistische modellering van onzekerheid moeten vastleggen.

Veel repliceerbaarheidsdiscussies vinden hun oorsprong in onenigheid over dit laatste. Wat voor de ene onderzoeker een 'replicatie' is, is dat niet altijd voor een andere (Gilbert, King, Pettigrew, & Wilson, 2016; Mathur, & VanderWeele, 2019; Maxwell, Lau, & Howard, 2015; Patil, Peng, & Leek, 2016). In de replicatiestudies van de Open Science Collaboration (2015) werd bijvoorbeeld met vier criteria gewerkt:

(1) Is er in de replicatiestudie een statistisch significant effect op het 5% significantieniveau in dezelfde richting als in de eerdere studie?
(2) Ligt de puntschatting voor de grootte van het effect van de eerdere studie binnen het 95% betrouwbaarheidsinterval van de replicatiestudie?
(3) Resulteert de combinatie van de informatie in de eerdere studie en de replicatiestudie in een statistisch significant effect?
(4) Wat is de subjectieve evaluatie van de replicatieteams?

Er zijn ondertussen echter vele alternatieve methoden en criteria voorgesteld (Ly, Etz, Marsman, & Wagenmakers, 2018; Simonsohn, 2015; Tackett & McShane, 2018).

Recent reikten Hedges en Schauer (2019a, 2019b, 2019c) een interessant perspectief op de repliceerbaarheidsdiscussie aan. Vanuit een meta-analytisch perspectief toonden zij aan dat één enkele replicatiestudie zelden voldoende onderscheidingsvermogen heeft om tot een eenduidige conclusie over replicatie te besluiten. Er is volgens hen een 'replicatie-programma' nodig met meerdere geplande en gepreregistreerde directe replicatiestudies (zie onder 5, aanbevelingen 1 en 2). Zelfs voor een dergelijk replicatieprogramma zijn er nog verschillende meta-analytische toetsen mogelijk, afhankelijk van het feit of de bewijslast bij replicatie of niet-replicatie wordt gelegd, of de replicatiestudies als identiek worden verondersteld dan wel of er een geringe mate van heterogeniteit wordt toegelaten, en of de replicatiestudies als de volledige populatie worden beschouwd of als een steekproef uit de populatie van mogelijk relevante replicatiestudies.

Het bestaan van deze verschillende perspectieven, van uiteenlopende operationaliseringen en criteria om consistentie bij replicaties vast te stellen – inclusief het gebruik van verschillende statistische modellen om de resultaten van replicatiestudies inzichtelijk te maken – hypothekeert het replicatieonderzoek geenszins. Deze diversiteit aan methoden en technieken is eigen aan wetenschappelijk onderzoek en maakt er ook de rijkdom van uit (Zwaan, Etz, Lucas, & Donnellan, 2018). Ook in het replicatieonderzoek zal het de kunst zijn *common ground* te vinden: ofwel door op voorhand en op basis van consensus goede afspraken te maken en de replicatiestudies te preregistreren, ofwel door een veelheid van methoden en technieken toe te passen en de convergenties en divergenties in kaart te brengen.

Kanttekeningen bij de definitie van 'repliceerbaarheid'

De term 'repliceerbaarheid' wordt in twee verwante betekenissen gebruikt. In dit Standpunt volgen we deze tweeledigheid, waarbij uit de context moet blijken over welke betekenis het gaat. De eerste betekenis wordt benadrukt door Peels (2019) en verwijst naar de loutere mogelijkheidsvoorwaarden om een replicatiestudie uit te voeren. Het gaat dan over de verzameling van eigenschappen waaraan

een eerdere studie moet voldoen om tot een replicatiestudie te *kunnen* leiden. In deze betekenis gaat 'repliceerbaarheid' dus over de kwaliteitskenmerken van de eerdere studie, zoals rapporteringskwaliteit, de beschikbaarheid van materiaal en de transparantie ervan. Dat zijn kenmerken die bij de aanbevelingen (zie Hoofdstuk 5) aan bod zullen komen.

De algemenere betekenis van 'repliceerbaarheid' wordt gehanteerd door de National Academies of Sciences, Engineering, and Medicine (2019) en beperkt zich niet tot een individuele studie. Ze verwijst naar de mogelijkheid van replicaties of naar de aanwezigheid van feitelijke replicaties in een bepaald onderzoeksdomein. Het is in deze betekenis dat repliceerbaarheid wordt beschouwd als een hoeksteen van wetenschappelijk kennis en als toetssteen voor de externe betrouwbaarheid en geldigheid van wetenschappelijke uitspraken (zie ook Earp & Trafimow, 2015; National Academies of Sciences, Engineering, & Medicine, 2019; Onghena, 1998; Open Science Collaboration, 2015; Srivastava, 2018; Wiersma, 1995; Zwaan et al., 2018).

De functie en de centrale plaats die repliceerbaarheid binnen het wetenschappelijk onderzoek krijgt, zijn niet zonder controverse. Repliceerbaarheid wordt namelijk geassocieerd met het logisch positivisme van Carnap (1928/1967) of met het kritisch rationalisme van Popper (1935/2007), epistemologische posities die door vele wetenschapsfilosofen na Kuhn (1962), Lakatos en Musgrave (1970) en Feyerabend (1975) zijn verlaten (Kwa, 2018).

In de wetenschappelijke methodologie is het werk van met name Karl Popper nog heel invloedrijk. Repliceerbaarheid staat daarin centraal: het falsificeren[3] of corroboreren van wetenschappelijke uitspraken is onmogelijk zonder het principe van repliceerbaarheid te aanvaarden. Met de woorden van Popper (1935/2007): 'The scientifically significant physical effect may be defined as that which can be regularly reproduced by anyone who carries out the appropriate experiment in the way prescribed.' (pp. 23-24). Een deel van de blijvende aantrekkingskracht van het Popperiaans gedachtegoed is te begrijpen vanuit de toegankelijkheid en het operationele karakter van de teksten van Popper: het gaat over de wetenschappelijke werkwijze en over procedures. Ook statistici begrijpen die taal, en de mengeling met de onzekerheid waarmee onderzoekers moeten afrekenen als ze metingen

[3] Gegeven het belang van de term 'falsificatie' in de wetenschapsmethodologie en -filosofie enerzijds en de mogelijke koppeling van repliceerbaarheidsproblemen aan *questionable research practices* en fraude anderzijds, is het verwarrend om het opzettelijk vervalsen van gegevens ook met de term 'falsificatie' aan te duiden (zie bv. Overbeke, 1994; Van Liedekerke, Van Driessche, & Nollet, 2019). Om deze verwarring te vermijden gebruiken we in dit Standpunt de term 'vervalsing' als we verwijzen naar het valselijk veranderen van verzamelde gegevens in empirisch onderzoek, en de term 'weerlegging' als we verwijzen naar de verwerping van wetenschappelijke theorieën en hypothesen.

verrichten en nieuwe steekproeven trekken, levert een krachtige cocktail op: 'In order to assert that a natural phenomenon is experimentally demonstrable we need, not an isolated record, but a reliable method of procedure. In relation to the test of significance, we may say that a phenomenon is experimentally demonstrable when we know how to conduct an experiment which will rarely fail to give us a statistically significant result.' (Fisher, 1935, p. 14)

Tegenover het falsificationisme van Popper wordt weleens de Duhem-Quinestelling geplaatst, die stelt dat cruciale experimenten – *experimentum crucis*, in de betekenis die Francis Bacon en Isaac Newton eraan gaven – niet bestaan: het is onmogelijk om met een experiment een individuele hypothese te weerleggen (Gillies, 1998; Harding, 1976). Omdat deze stelling botst met de gangbare wetenschappelijke methodologie en wijst op een fundamenteel probleem bij empirisch onderzoek, maken we even een excursie over de rol van dit probleem bij repliceerbaarheid.

De Duhem-Quinestelling heeft als uitgangspunt dat uit een hypothese geen rechtstreekse operationaliseringen en predicties volgen. Die operationaliseringen en predicties kunnen alleen maar worden afgeleid door *bijkomende veronderstellingen* over de wereld te maken. In replicatiestudies gaat het dan bijvoorbeeld over de betrouwbaarheid van de meetinstrumenten of de irrelevantie van bepaalde contextuele factoren. Als de resultaten van een studie vervolgens niet in overeenstemming zijn met de predicties vanuit een hypothese, is het altijd mogelijk om de hypothese tegen weerlegging te beschermen door de bijkomende veronderstellingen als verklaring in te roepen: het instrument was toch niet zo betrouwbaar als verondersteld, of het effect treedt alleen bij een bepaalde subgroep van personen op of iets dergelijks. De hypothese en de bijkomende veronderstellingen vormen in die optiek een *Gestalt* (een zogenaamde *bundle of hypotheses*) waaruit één specifieke onderzoekshypothese onmogelijk te isoleren is. De *bundle of hypotheses* kan worden getoetst en kan in zijn geheel worden weerlegd als de resultaten van een studie niet met de predicties in overeenstemming zijn, maar een rechtstreekse toets van één hypothese afgeleid uit de theorie is uitgesloten.

Voor de repliceerbaarheidsdiscussie is de Duhem-Quinestelling relevant omdat ze waarschuwt voor een te naïeve kijk op replicatiestudies. De stelling – in haar meest radicale vorm: de onmogelijkheid om afzonderlijke hypothesen te toetsen – lijkt echter weinig vruchtbaar voor de wetenschappelijke praktijk. Ze lijkt eerder een uitspraak te zijn over de onmogelijkheid van strikt *kennistheoretische zekerheid* op basis van empirische toetsen, waarbij abstractie wordt gemaakt van de menselijke en sociale kant van wetenschap (het wetenschappelijk forum, het debat en de consensusvorming) en van de onzekerheid die wetenschappelijke uitspraken blijvend kenmerken (Leezenberg & de Vries, 2017).

Bovendien lijkt de Duhem-Quinestelling meer te gaan over de onoverbrugbare kloof tussen theorie en empirie dan over een vergelijking van empirische toetsen onderling. Stel dat iemand beweert dat psychokinese (het verplaatsen of beïnvloeden van materiële objecten louter en alleen op basis van gedachten en wilskracht) op experimenteel overtuigende wijze is aangetoond (Bösch, Steinkamp, & Boller, 2006). Het volstaat dan om het experiment zo zorgvuldig mogelijk te herhalen, en eventueel nog eens en nog eens, om er de repliceerbaarheid van na te gaan. De stap naar theorievorming of de implicaties van de al dan niet repliceerbaarheid van het experiment voor een bestaande theorie over de menselijke geest is finaal uitermate belangrijk, maar is in eerste instantie niet wezenlijk voor het repliceerbaarheidsdebat.

Als tentatieve conclusie kunnen we stellen dat repliceerbaarheid een belangrijk kenmerk van wetenschappelijke uitspraken blijft, ondanks de fundamentele vragen. Zonder repliceerbaarheid zou er ook geen *evidence-based* praktijk mogelijk zijn; er zou in dat geval geen enkele garantie zijn dat bijvoorbeeld de door onderzoek aangetoonde effectiviteit van cognitieve gedragstherapie bij een majeure depressie (Gartlehner et al., 2016) ook in de toekomst of voor andere personen dan de personen uit het onderzoek geldig blijft. De fundamentele vragen maken wél duidelijk dat in dit debat wetenschapsfilosofie een plaats heeft (Morawski, 2019). We komen hierop terug bij het formuleren van de aanbevelingen (zie aanbeveling 5 in Hoofdstuk 5).

Nog een laatste kanttekening

In de titel van dit Standpunt hebben we de kwalificatie 'crisis' weggelaten. We willen het over 'repliceerbaarheid' hebben, niet louter over het crisisaspect ervan. Door deze woordkeuze vermijden we dat we moeten omschrijven wat een crisis is en of we met het repliceerbaarheidsprobleem in de empirische wetenschappen al dan niet de noodtoestand moeten uitroepen (Jamieson, 2018; Nelson, Simmons, & Simonsohn, 2018; Pashler, & Harris, 2012).

De kwalificatie 'crisis' is overigens een allesbehalve neutrale beschrijving van de stand van zaken van repliceerbaarheid in de empirische menswetenschappen. Ze houdt in dat we te maken zouden hebben met een uitzonderlijk dramatisch fenomeen en een ongeziene omwenteling, en roept op tot daadkrachtige actie om de crisis het hoofd te bieden (Maxwell et al., 2015; Morawski, 2019; Stroebe & Strack, 2014). Pashler en Harris (2012) merken terecht op dat er enige scepsis nodig is als de zoveelste zogenaamde 'crisis' uitbreekt. Ondertussen hebben we naast de 'repliceerbaarheidscrisis' immers ook al een 'theoriecrisis' (Oberauer & Lewandowsky, 2019), een 'meetcrisis' (Flake & Fried, 2019; Meier, 1994; Schimmack, 2019) en een 'crisis dat we voortdurend in crisissen zitten' (Hughes, 2018).

Het gebruik van het woord 'crisis' kan een retorische truc worden om meer aandacht voor een probleem te vragen. Dat is volgens ons reden genoeg om zuinig met het woord om te gaan. Met dit Standpunt willen we daarom niet alarmistisch doen, maar wel proberen om zo objectief mogelijk de beschikbare repliceerbaarheidsresultaten in de empirische menswetenschappen in kaart te brengen en om, voor het Vlaamse menswetenschappelijk onderzoek, waar nodig en gepast, aanbevelingen tot actie te formuleren. In dat opzicht sluit dit Standpunt aan bij eerdere aanbevelingen vanuit The Academy of Medical Sciences in het Verenigd Koninkrijk (2015), de Koninklijke Nederlandse Akademie van Wetenschappen (2018), en de National Academies of Sciences, Engineering, and Medicine in de Verenigde Staten (2019).

2.2.2. Aanverwante begrippen: reproduceerbaarheid, robuustheid en generaliseerbaarheid

Bij discussies over repliceerbaarheid worden soms begrippen betrokken die wij niet als de focus van dit Standpunt beschouwen. Het gaat met name over 'reproduceerbaarheid', 'robuustheid' en 'generaliseerbaarheid'. Een definiëring hiervan stelt ons in staat om het begrip 'repliceerbaarheid' nog beter af te bakenen:

> *Reproduceerbaarheid*
>
> Reproduceerbaarheid verwijst naar een eigenschap van een studie waarbij het gebruik van het oorspronkelijke gegevensbestand van die studie en van dezelfde analysetechnieken identieke resultaten oplevert als in de oorspronkelijke studie of de rapportering ervan (Bollen, Cacioppo, Kaplan, Knosnick, & Olds, 2015; National Academies of Sciences, Engineering, & Medicine, 2019).

Plesser (2018) merkt op dat er bij het gebruik van de termen 'repliceerbaarheid' en 'reproduceerbaarheid' in verschillende wetenschappelijke disciplines verschillende tradities bestaan, wat tot spraakverwarring kan leiden. Een van de eerste grootschalige replicatieprogramma's in de psychologie had bijvoorbeeld als titel *Reproducibility Project: Psychology* (Open Science Collaboration, 2015), terwijl het manifest over repliceerbaarheid en niet over reproduceerbaarheid ging. In latere publicaties over gelijkaardige projecten wordt er echter wel steeds over '*replicability*' gesproken (Camerer et al., 2016, 2018; Ebersole et al., 2016; Klein, 2014, 2018, 2019).

Aan reproduceerbaarheid wordt veel belang gehecht in de technische wetenschappen, in het bijzonder in de computerwetenschappen. Reproduceerbaarheid lijkt als voorwaarde voor de geldigheid van wetenschappelijke uitspraken vanzelfsprekend en weinig problematisch, maar is dat niet, zeker niet in de computerwetenschappen. De eigenschap verwijst immers naar *identieke* resultaten. Bij grootschalige

berekeningen, complexe algoritmen, stapsgewijze procedures en simulaties blijkt het niet altijd eenvoudig om de resultaten exact te reproduceren of om ze op een reproduceerbare manier te rapporteren. Er is de problematiek van, onder meer, de afrondings- en simulatiefouten, de eindige nauwkeurigheid van computers, de instabiliteit van complexe algoritmen, verschillen tussen (verschillende versies van) *operating systems* en infrastructuurafhankelijkheid (Elmenreich, Moll, Theuermann, & Lux, 2018, Plesser, 2018). De helft van het ongeveer 200 bladzijden tellende rapport over *Reproducibility and replicability in science* (National Academies of Sciences, Engineering, and Medicine, 2019) gaat trouwens over reproduceerbaarheid.

Voor de empirische menswetenschappen is reproduceerbaarheid niet de grootste zorg. Hier kan worden volstaan met op te merken dat van analyseresultaten die niet herhaalbaar zijn met hetzelfde gegevensbestand (reproduceerbaarheid), *a fortiori* niet kan worden verwacht dat ze met een nieuw gegevensbestand herhaalbaar zullen zijn (repliceerbaarheid). In die zin is reproduceerbaarheid een voorwaarde om tot repliceerbaarheid te kunnen komen. Voor de empirische menswetenschappen is voornamelijk de correcte en volledige documentatie van de statistische analyse van belang (Nuijten, Bakker, Maassen, & Wicherts, 2018; Nuijten, Hartgerink, van Assen, Epskamp, & Wicherts, 2016).

Robuustheid

Robuustheid verwijst naar een eigenschap van een studie waarbij het gebruik van het oorspronkelijke gegevensbestand van die studie en van *andere* analysetechnieken vergelijkbare resultaten oplevert als in de oorspronkelijke studie of de rapportering ervan (Box, 1953; Box, Leonard, & Wu, 1983; Hansen & Sargent, 2008).

Net zoals bij reproduceerbaarheid gaat het bij robuustheid over hetzelfde gegevensbestand, maar bij robuustheidsonderzoek worden de analysetechnieken gevarieerd om na te gaan of de resultaten gevoelig zijn voor het gekozen soort analysetechniek. Dikwijls gaat het over het vergelijken van de veronderstellingen die aan de analysetechnieken ten grondslag liggen (Hansen & Sargent, 2008). Robuustheidsonderzoek is minder rechtstreeks relevant voor het repliceerbaarheidsvraagstuk als de methoden van de replicatiestudie een exacte replica van de methoden van de eerdere studie zijn.

Generaliseerbaarheid

Generaliseerbaarheid verwijst naar een eigenschap van een studie waarbij het verzamelen van nieuwe gegevens en het gebruik van andere methoden vergelijkbare resultaten oplevert als in de oorspronkelijke studie of de rapportering ervan (Cronbach, 1982; Shadish, Cook, & Campbell, 2002).

Generaliseerbaarheid gaat over de veralgemeenbaarheid van de resultaten van de studie naar bijvoorbeeld andere populaties, situaties en tijdstippen. Sommige vormen van conceptuele replicatie kunnen als onderzoek naar de generaliseerbaarheid worden beschouwd (LeBel et al., 2018). Generaliseerbaarheid is veel moeilijker te bereiken dan repliceerbaarheid. In de omschrijving hierboven gaat de repliceerbaarheidsvraag vooraf aan de generaliseerbaarheidsvraag.

Tabel 1 brengt bovenstaande drie aanverwante begrippen samen met het begrip 'repliceerbaarheid' om de belangrijkste verschilpunten te benadrukken (Schloss, 2018). In de tabel worden slechts twee dichotomieën met elkaar gekruist: 'dezelfde versus verschillende methode' en 'dezelfde versus verschillende data'. De tabel houdt uiteraard een begripsvereenvoudiging in, maar stelt de verschilpunten wel op scherp. De vier begrippen verschijnen elk als een specifieke combinatie van de twee dichotomieën.

Repliceerbaarheid wordt in het schema voorgesteld als 'dezelfde methode, verschillende data' en kan als een sleutelkenmerk van *externe betrouwbaarheid* worden beschouwd (Onghena, 1998; Wiersma, 1995). Reproduceerbaarheid verwijst dan naar *interne betrouwbaarheid*, en robuustheid en generaliseerbaarheid naar respectievelijk *interne* en *externe validiteit*. Bovendien bestaat er een afhankelijkheid tussen de vier begrippen. Betrouwbaarheid kan als een voorwaarde voor validiteit worden beschouwd. Het heeft weinig zin om na te gaan of vergelijkbare resultaten worden bekomen met andere methoden als dezelfde methode nog niet tot vergelijkbare resultaten leidt. Op dezelfde manier kan men het interne kwaliteitscriterium als een voorwaarde voor het externe kwaliteitscriterium beschouwen. Het heeft geen zin om na te gaan of vergelijkbare resultaten worden bekomen met andere data als het gebruik van dezelfde data nog niet tot vergelijkbare resultaten leidt.

Met het schema van tabel 1 in het achterhoofd zullen we in de volgende paragrafen het beschikbare replicatieonderzoek presenteren, de implicaties ervan bespreken en tot slot aanbevelingen formuleren.

Tabel 1

Rooster met de begrippen die worden gehanteerd om de betrouwbaarheid en geldigheid van wetenschappelijke resultaten te beschrijven (Schloss, 2018)

	Dezelfde data	**Verschillende data**
Dezelfde methode	Reproduceerbaarheid	Repliceerbaarheid
Verschillende methode	Robuustheid	Generaliseerbaarheid

3. Antecedenten, stand van zaken en mogelijke oorzaken

3.1. Antecedenten

Het repliceerbaarheidsprobleem heeft een lange voorgeschiedenis (zie bv. Bower & Mayer, 1985; Collins, 1985; Shapin & Schaffer, 1985; Smith, 1970) maar werd in 2005 op zijn scherpst geformuleerd in een publicatie van Ioannidis met de geruchtmakende titel *Why Most Published Research Results Are False*. Daarin simuleerde Ioannidis (2005b) gegevens met een aantal scenario's die het huidige wetenschappelijk onderzoek kenmerken. De simulaties toonden aan dat voor de meeste designs en settings de kans op een valse onderzoeksbevinding groter was dan die op een ware onderzoeksbevinding. Bovendien bleek uit de simulaties dat de kans op een ware onderzoeksbevinding kleiner is bij een lager statistisch onderscheidingsvermogen (m.a.w. indien het gaat over een studie met een kleiner aantal deelnemers en/of indien de effectgroottes kleiner zijn), indien meerdere studies dezelfde onderzoeksvraag hebben, indien er meerdere *researcher degrees of freedom* zijn (d.i. een grotere flexibiliteit in designs, operationele definities, uitkomstmaten en statistische technieken waaruit de onderzoeker een keuze kan maken), indien er grotere *conflicts of interest* en meer vooroordelen zijn en indien er binnen een bepaald domein meer onderzoeksteams statistische significantie najagen. Hoewel het in deze methodologische exploratie van Ioannidis (2005b) over gesimuleerde gegevens ging, waren de scenario's voor vele onderzoekers herkenbaar. De conclusie sloeg dan ook in als een bom: 'Claimed research findings may often be simply accurate measures of the prevailing bias.' (p. 700)

De impact van de bom werd nog groter toen Ioannidis (2005a) in datzelfde jaar met gegevens van daadwerkelijk uitgevoerd onderzoek aan de slag ging. Hij voerde een systematische review uit waarin hij naging op welke manier alle originele klinische studies die waren gepubliceerd in gerenommeerde medische tijdschriften tussen 1990 en 2003 en die meer dan duizend keer in de wetenschappelijke literatuur werden geciteerd, in de daaropvolgende jaren werden gerepliceerd. Als er replicatiestudies met een vergelijkbare of grotere steekproef of een vergelijkbaar of verbeterd design waren, vergeleek hij de resultaten van de replicatiestudie met die van de oorspronkelijke studie.

Weinigen hadden het resultaat zien aankomen. Van de 34 oorspronkelijke studies waarvoor replicatiestudies bestonden, hielden er slechts 20 (59%) stand bij replicatie. Bij 7 (20.5%) studies was het effect in de replicatiestudie beduidend kleiner dan in de oorspronkelijke studie en bij nog 7 (20.5%) andere kreeg je zelfs een effect dat volledig in de andere richting ging. Met de simulatiestudie van Ioannidis (2005b) in het achterhoofd zijn deze resultaten weinig verrassend, en misschien nog positiever dan kon worden verwacht. In de systematische review waren er echter ook nog 4 originele studies met negatieve resultaten en 11 veel geciteerde studies met positieve resultaten waarvoor in de wetenschappelijke

literatuur nog geen replicatiestudie kon worden teruggevonden. Als deze studies worden meegeteld, is de uiteindelijke replicatiebalans minder positief (20/49 = 41%).

Een aantal jaar later verscheen vanuit neurowetenschappelijke hoek de *voodoo correlations*-paper van Vul, Harris, Winkielman en Pashler (2009). Daarin maakten de auteurs heel tastbaar wat het betekent om statistische significantie na te jagen terwijl er een grote flexibiliteit in analysemogelijkheden bestaat. Zij toonden aan dat de extreem hoge correlaties tussen hersenactiviteit (gemeten via *functional magnetic resonance imaging*, fMRI) en persoonlijkheidskenmerken die in vele neurowetenschappelijke studies werden aangetroffen, het resultaat zijn van een ongerapporteerde filtering en voorbewerking van de data. In een enquête bij de auteurs van 55 fMRI-publicaties waarin dergelijke extreem hoge correlaties werden gevonden, maar waarbij de details over de filtering en voorbewerking in de publicaties ontbraken, gaf meer dan de helft van de auteurs toe dat ze een gangbare analysetechniek volgden die bij nader toezien tot een misleidende inflatie van correlaties kan leiden. Bovendien leidde de techniek in een aantal publicaties tot volstrekte *spurious* ('onechte') correlaties: die zijn niet als inhoudelijke samenhang tussen twee verschijnselen interpreteerbaar, maar wel volledig te verklaren vanuit een samenhang met andere verschijnselen.

Bennett, Baird, Miller en Wolford (2009) deden er nog een schepje bovenop door aan een dode zalm 15 foto's van mensen in sociale en emotioneel beladen situaties te tonen. Aan de zalm werd gevraagd om in te schatten welke emotie de persoon op de foto ervoer. Door gebruik te maken van gesofisticeerde fMRI-analyses konden Bennett et al. (2009) aantonen dat er verschillende regionen in de zalmhersenen statistisch significant actiever waren tijdens de fotosessie dan tijdens de rustsessies daartussenin (voor verdere duiding zie ook Bennett, Miller en Wolford, 2009).

Het laat zich raden dat het niet lang duurde alvorens ook vanuit de empirische menswetenschappen vergelijkbare onrustwekkende signalen kwamen. Er waren met name twee wake-upcalls. In een publicatie van Daryl Bem (2011) werd over negen experimenten gerapporteerd die aantoonden dat mensen beïnvloed kunnen worden door een onvoorspelbare gebeurtenis in de toekomst, een resultaat dat achteraf onrepliceerbaar bleek (Galak, LeBoeuf, Nelson, & Simmons, 2012; Ritchie, Wiseman, & French, 2012). Er was ook Diederik Stapels bekentenis van grootschalige gegevensvervalsing na een opzienbarend eindrapport van drie onafhankelijke onderzoekscommissies (Levelt, Drenth, & Noort, 2012). Wakker geschud door die twee publicaties verscheen dan in 2011 een artikel waarvan de titel boekdelen spreekt: *False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant* (Simmons, Nelson, & Simonsohn, 2011). Daarin rapporteerden de auteurs onder andere een studie waarin ze twintig universiteitsstudenten lieten luisteren naar de song *When*

I'm Sixty-Four van The Beatles. Ze verzamelden bij hen ook een veelheid aan informatie die als controlevariabelen bij de statistische analyse konden worden betrokken. Zonder gegevens te moeten vervalsen, maar met een creatieve en selectieve statistische analyse, demonstreerden ze hoe eenvoudig het is om evidentie te vinden voor de stelling dat het luisteren naar *When I'm Sixty-Four* ervoor zorgt dat de chronologische leeftijd van de studenten afneemt. Met hun studie traden ze dus in de voetsporen van Ioannidis (2005b) en Bennett et al. (2009), die eerder aantoonden dat een groot aantal *researcher degrees of freedom* kan leiden tot absurde resultaten en/of interpretaties.

De *mogelijkheid* tot absurde resultaten en/of interpretaties zegt nog niets over de daadwerkelijke aanwezigheid van dergelijke vals positieven in de wetenschappelijke literatuur. Zoals vermeld, voerde Ioannidis (2005a) een systematische review uit om vals positieven op te sporen, maar de methode die hij volgde, kan het probleem van de vals positieven onderschatten omdat in zijn systematische review alleen gepubliceerde replicatiestudies waren opgenomen. *Editors* en *reviewers* van toonaangevende tijdschriften zouden terughoudend kunnen zijn om negatieve replicatiestudies over eerdere 'doorbraakartikelen' te publiceren.[4] Bovendien bleek al uit de *review* van Ioannidis (2005a) dat van 11 van de 49 originele studies (22%) geen enkele replicatiestudie in de wetenschappelijke literatuur terug te vinden was.

Twee alternatieve sporen die kunnen worden gevolgd, zijn (1) nagaan of er aanwijzingen voor 'bedenkelijke' onderzoekspraktijken in de oorspronkelijke studies terug te vinden zijn en (2) de replicatiestudies zelf uitvoeren. Het eerste spoor betreft het onderzoek naar de zogenaamde *questionable research practices* (QRPs), waarbij de veronderstelling is dat de QRPs de kans vergroten dat de resultaten van een studie niet repliceerbaar zijn (John, Loewenstein, & Prelec, 2012). Dit bespreken we in sectie 3.3. Het tweede spoor betreft de grootschalige replicatieprojecten die onder impuls van Brian Nosek in verschillende wetenschapsdomeinen werden opgezet (Open Science Collaboration, 2015). Hierop gaan we in de volgende sectie in.

[4] Er is overigens wel wat evidentie voor een dergelijke terughoudendheid. De auteurs van de oorspronkelijke studies worden niet zelden door *editors* als (anonieme) *reviewers* van replicatiestudies uitgenodigd. Als in de replicatiestudie een kritische houding tegenover de oorspronkelijke studie wordt aangenomen, komt het voor dat de auteurs van de oorspronkelijke studie strenger oordelen over methodologische tekortkomingen, verklaringen zoeken en vinden voor het eventuele kleinere of afwezige effect, of de studie diskwalificeren als niet origineel genoeg. Dit kan op zijn beurt leiden tot zelfselectie en -censuur bij onderzoekers om zich met replicatiestudies in te laten (Bronstein, 1990; French, 2012; Martin & Clarke, 2017; Neuliep & Crandall, 1990, 1993; Schmidt, 2009).

3.2. Stand van zaken

Het rapport over het eerste grootschalige replicatieproject in de empirische menswetenschappen verscheen in 2015 in het invloedrijke tijdschrift *Science* (Open Science Collaboration, 2015). Het is om meerdere redenen opmerkelijk. Niet alleen gaat het over een indrukwekkende verzameling van zorgvuldig opgezette replicatiestudies en zijn de resultaten opzienbarend, maar het is ook een van de eerste rapporten in dit genre dat gebruik maakt van *crowdsourced science* (Uhlmann et al., 2019). In een breed opgezette samenwerking tussen internationaal verspreide laboratoria werden in dat project honderd replicatiestudies uitgevoerd. Om de honderd oorspronkelijke studies te selecteren werd een steekproefkader van publicaties uit 2008 in drie belangrijke psychologietijdschriften gehanteerd: *Psychological Science*, *Journal of Personality and Social Psychology* en *Journal of Experimental Psychology: Learning, Memory, and Cognition*.

Bij de bespreking van de resultaten geven de auteurs vooraf een belangrijke waarschuwing mee: er bestaat geen eenduidig criterium om replicatiesucces te definiëren. Zoals we in paragraaf 2.2.1 al aanstipten, werd er daarom voor gekozen om met meerdere operationaliseringen te werken. We herhalen ze even:

(1) Is er in de replicatiestudie een statistisch significant effect op het 5% significantieniveau in dezelfde richting als in de eerdere studie?
(2) Ligt de puntschatting voor de grootte van het effect van de eerdere studie binnen het 95% betrouwbaarheidsinterval van de replicatiestudie?
(3) Resulteert de combinatie van de informatie in de eerdere studie en de replicatiestudie in een statistisch significant effect?
(4) Wat is de subjectieve evaluatie van de replicatieteams?

Wat de eerste operationalisering betreft, werd slechts in 36% van de replicatiestudies een statistisch significant effect in de dezelfde richting als de oorspronkelijke studie gedetecteerd (ondanks een aanzienlijk onderscheidingsvermogen en zorgvuldige controle en metingen). Wat de tweede operationalisering betreft, kon de gemiddelde effectgrootte op ongeveer 50% van de gemiddelde effectgrootte van de oorspronkelijke studies worden geschat. Slechts 47% van de puntschattingen voor de effectgrootte van de oorspronkelijke studie lag binnen het 95% betrouwbaarheidsinterval van de replicatiestudie. Wat de derde operationalisering betreft, resulteerde de combinatie van informatie in de eerdere studie en de replicatiestudie tot 68% statistisch significante effecten. De vierde operationalisering ten slotte gaf aan dat de replicatieteams slechts 39% van de effecten als gerepliceerd beschouwden. Deze percentages voor replicatiesucces in psychologisch onderzoek van minder dan 40% (tot 68%, afhankelijk van de operationalisering) zijn opvallend laag en hun orde van grootte is vergelijkbaar met de 41% en 59% replicatiesucces voor klinische studies zoals kon worden afgeleid uit de review van Ioannidis (2005a).

Een wisselend replicatiesucces werd ook geobserveerd in het Many Labs-project bij een reeks *crowdsourced* studies met een hoog onderscheidingsvermogen (Ebersole et al., 2016; Klein et al., 2014, 2018, 2019). De klemtoon van dit project lag op psychologisch onderzoek, net zoals bij het project van de Open Science Collaboration (2015). Zonder in te gaan op de details van deze replicatiestudies werden percentages replicatiesucces opgetekend van 77% (10/13 in Many Labs 1, Klein et al., 2014), 50% (14/28 in Many Labs 2, Klein et al., 2018), 30% (3/10 in Many Labs 3, Ebersole et al., 2016) en 5% (1/21 in Many Labs 4, Klein et al. 2019). De variatie in de percentages werd voornamelijk toegeschreven aan de geselecteerde effecten, niet aan de gekozen deelnemers, de setting, de timing of de betrokkenheid van de oorspronkelijke auteurs (Ebersole et al., 2016; Klein et al., 2014, 2018, 2019).

Hoewel het gros van de replicatiestudies op het domein van de psychologie plaatsvindt, zijn er ook al eerste resultaten van grootschalige replicatieprojecten in andere empirische menswetenschappen. Op het domein van de *experimentele economie* rapporteerden Camerer et al. (2016) in *Science* de resultaten van 18 replicatiestudies van experimenten die tussen 2011 en 2014 in de *American Economic Review* en de *Quarterly Journal of Economics* zijn verschenen. Zij vonden bij 11 van de 18 replicatiestudies (61%) statistisch significante effecten die in dezelfde richting gingen als in de oorspronkelijke studie. Bovendien bedroeg het gemiddelde effect in de replicatiestudies slechts 66% van het oorspronkelijke. Naast andere operationaliseringen van replicatiesucces maakten zij als een van de eersten uitvoerig gebruik van een *prediction market*-maat. In een dergelijke *prediction market* kunnen onderzoekers die vertrouwd zijn met het betreffende onderzoek, aandelen voor een bepaalde studie kopen of verkopen, waarbij de geldwaarde van het aandeel afhankelijk is van het verwachte replicatiesucces. Deze *prediction markets* genereren op die manier een collectieve 'marktreplicatiekans' die als een operationalisering van het verwachte replicatiesucces kan worden geïnterpreteerd. Camerer et al. (2016) vonden een gemiddelde marktreplicatiekans van 75.2%, 10% hoger dan het daadwerkelijke replicatiesucces (maar niet statistisch significant verschillend). Bovendien was er een positieve rangcorrelatie tussen de marktreplicatiekans en de mate van replicatiesucces, maar ook die was niet statistisch significant verschillend van 0.

Op het domein van de *sociale wetenschappen* vonden Camerer et al. (2018) bij 13 van de 21 replicatiestudies (62%) statistisch significante effecten in dezelfde richting als in de oorspronkelijke studies (gepubliceerd in *Nature* en *Science* tussen 2010 en 2015). Het gemiddelde effect in de replicatiestudies was 50% van het oorspronkelijke. Dat is lager dan het percentage bij economie (Camerer et al., 2016) en op hetzelfde niveau als bij psychologie (Open Science Collaboration, 2015). In de *prediction market* vonden Camerer et al. (2018) een gemiddelde marktreplicatiekans van 63.4%, wat vergelijkbaar is met het daadwerkelijke replicatiesucces. Bovendien werd er een hoog positieve en

statistisch significante rangcorrelatie van 0.842 tussen de marktreplicatiekans en de mate van replicatiesucces gevonden. Dit betekent dat het vanuit de specifieke onderzoeksgemeenschap mogelijk was om te voorspellen welke resultaten repliceerbaar zijn en bij welke resultaten er problemen te verwachten zijn.

Ook in de *politieke wetenschappen* (Franco, Malhotra, & Simonovits, 2014), de *pedagogische wetenschappen* (Makel & Plucker, 2014), de *didactiek* (Cooper, 2018), de *criminologie* (Pridemore, Makel, & Plucker, 2018), de *financiële wetenschappen* (Harvey, Liu, & Zhu, 2016) en de *micro- en macro-economische wetenschappen* (Christensen & Miguel, 2018; Ioannidis, Stanley, & Doucouliagos, 2017) werden de repliceerbaarheidsproblemen voor het voetlicht gebracht. Uit een enquête van *Nature* bij meer dan 1500 onderzoekers bleek bovendien dat die problemen niet beperkt blijven tot de empirische menswetenschappen (Baker, 2016): 87% van de chemici, 78% van de biologen, 69% van de natuurkundigen en ingenieurs, 68% van de medici en 67% van de geologen gaven aan dat ook zij al meemaakten dat het onmogelijk was om andermans experiment te repliceren.

3.3. Mogelijke oorzaken

Als mogelijke oorzaken van repliceerbaarheidsproblemen worden dikwijls de *questionable research practices* (QRPs) genoemd (John et al., 2012; Shrout & Rodgers, 2018; Nelson et al., 2018; Wicherts et al., 2016). Dat zijn praktijken die in bepaalde onderzoeksdomeinen of laboratoria gangbaar zijn en tot publicaties leiden, maar die methodologische of statistische problemen in zich dragen waardoor het risico groter wordt dat ze niet-repliceerbare resultaten bevatten (John et al., 2012).

QRPs moeten worden onderscheiden van regelrechte fraude, datafabricatie of datavervalsing. Hoewel wetenschapsfraude voorkomt en door repliceerbaarheids-problemen kan worden ontmaskerd, zijn QRPs subtieler en zijn de onderzoekers die QRPs hanteren zich vaak van geen kwaad bewust (Levelt et al., 2012; Miller & Hersen, 1992; Nelson et al., 2018). Door hun alomtegenwoordigheid zijn QRPs echter even desastreus voor de repliceerbaarheid als bewuste fraude (Agnoli, Wicherts, Veldkamp, Albiero, & Cubelli, 2017; John et al., 2012; Wicherts et al., 2016).

In de methodologische literatuur zijn vele QRPs beschreven. Wij onderscheiden hieronder drie types: (1) QRPs die te maken hebben met de onbetrouwbaarheid van de proefopzet, de dataverzameling en de data-analyse, (2) QRPs die te maken hebben met vertekening in de rapportering, en (3) QRPs die te maken hebben met selectiviteit in de rapportering.

Omdat repliceerbaarheid als een aspect van externe betrouwbaarheid kan worden beschouwd (zie paragraaf 2.2.2), worden de oorzaken bij het eerste type van QRPs

gezocht bij de *betrouwbaarheid* in de planning en uitvoering van de oorspronkelijke studie. Mogelijks zijn er ook problemen met de validiteit en de koppeling tussen het empirische en het theoretische niveau (zie bv. Borghi & Fini, 2019; Klein, 2014; Oberauer & Lewandowsky, 2019), maar directe repliceerbaarheid betreft in eerste instantie een voorafgaande stap, namelijk de vaststelling (het bestaan) en de herhaalbaarheid van een bepaald empirisch fenomeen. We onderscheiden, naast dit betrouwbaarheidsprobleem, bovendien twee types van QRPs die met rapportering te maken hebben. Deze dubbele aandacht voor rapportering ligt voor de hand, omdat herhaalbaarheid in grote mate afhankelijk is van een nauwkeurige en volledige verslaggeving over de oorspronkelijke studie. Als bepaalde details van de oorspronkelijke studie vertekend of onvolledig worden weergegeven, kunnen we niet verwachten dat andere onafhankelijke onderzoekers de studie kunnen herhalen.[5]

De zoektocht naar mogelijke oorzaken van het repliceerbaarheidsprobleem is belangwekkend omdat de identificatie van een oorzaak het probleem kan helpen oplossen door die oorzaak te verwijderen. Merk op dat de QRPs zijn geformuleerd op het niveau van het onderzoeksproces, wat echter niet wil zeggen dat we de verantwoordelijkheid voor het repliceerbaarheidsprobleem op het individuele niveau (bij de onderzoeker en het onderzoeksteam) leggen. Hoewel QRPs misschien wel als proximale oorzaken met de vinger kunnen worden gewezen, liggen oorzaken op vele andere niveaus aan de basis van de QRPs of houden ze die in stand (zie ook De Boeck & Jeon, 2018).

Ten eerste is er het publicatie-, communicatie- en mediabeleid. Als repliceerbaarheid te lijden heeft onder rapporteringsproblemen, dan spelen publicatiestrategieën, *peerreview*procedures en de redactionele politiek van wetenschappelijke tijdschriften zeker een rol bij het ontstaan en de instandhouding van repliceerbaarheidsproblemen (Ellemers, 2013; Hopf, Krief, Mehta, & Matlin, 2019; Lindsay, 2015). *Publicatiebias* is een prominent voorbeeld van een vertekening die tot repliceerbaarheidsproblemen kan leiden en die het individuele niveau overstijgt (Bishop, 2019). Dergelijke *publicatiebias* bij wetenschappelijke tijdschriften, in de richting van statistisch significante, originele, 'sexy' en onverwachte bevindingen, verhoogt het risico op vals positieven (Francis, 2012, 2014; Simmons et al. 2011).

Ten tweede is er het universitaire beleid. Als de aanwerving van onderzoekers en de promotie binnen het zelfstandig academisch personeel gekoppeld zijn aan criteria die hoofdzakelijk van publicaties in wetenschappelijke tijdschriften afhankelijk zijn, dan worden problemen door dat universitaire beleid op het publicatieniveau

[5] Bishop (2019) noemt als vier belangrijkste oorzaken van het repliceerbaarheidsprobleem: een laag onderscheidingsvermogen, HARKing (*Hypothesizing After the Results are Known*), *p-hacking* en *publication bias*. In onze typologie behoort een laag onderscheidingsvermogen tot het bredere probleem van de betrouwbaarheid van de proefopzet; HARKing is een voorbeeld van vertekend rapporteren en *p-hacking* en *publication bias* zijn voorbeelden van selectief rapporteren.

uitvergroot (Gunsalus & Robinson, 2018; Rodgers & Shrout, 2018; Smaldino & McElreath, 2016). Daartegenover staat dat universiteiten een belangrijke rol hebben als poortwachters voor zorgvuldige wetenschap, als voedingsbodem voor een onderzoekscultuur waarin repliceerbaarheid kan gedijen en als doorgeefluik voor betrouwbare onderzoeksmethoden aan toekomstige generaties onderzoekers via onderwijs, doctoraatsopleidingen, onderzoeksstages en permanente vorming (Artino, Driessen, & Maggio, 2019; Ayris, López de San Román, Maes, & Labastida, 2018; Krishna & Peter, 2018; Nosek et al., 2012, 2015).

Ten derde wordt het universitaire beleid gevoerd in een nationale en internationale politieke, maatschappelijke en financiële context die bepaalde types van onderzoek aanzien geeft en financieel aanmoedigt (Grimes, Bauch, & Ioannidis, 2018; Kiai, 2019; Lilienfeld & Waldman, 2017). Als universiteiten bijvoorbeeld deels worden gefinancierd in verhouding tot de hoeveelheid onderzoeksoutput (aantallen doctoraatsproefschriften en publicaties), dan kan men ook verwachten dat het universitaire onderzoeksbeleid zich deels op die kwantitatieve indicatoren gaat afstemmen. We merken hierbij op dat een weging van onderzoeksoutput in termen van 'kwaliteit' dikwijls weinig zoden aan de dijk zet als die kwaliteitsweging gemakshalve wordt gekwantificeerd als een *journal impact factor* (of afgeleide maten). Vanwege de grote financiële belangen en implicaties zullen uitgeverijen en tijdschriftredacties altijd proberen om handig op deze 'kwaliteitsindicatoren' in te spelen, bijvoorbeeld door reviewstudies en *special issues* te publiceren of door in te zetten op controverse, kortetermijnciteerbaarheid en spectaculaire bevindingen (Brito & Rodríguez-Navarro, 2019; Caon, 2017; Leydesdorff, Bornmann, Comins, & Milojevic, 2016; Seglen, 1997; The PLoS Medicine Editors, 2006). Dezelfde mechanismen interfereren ook bij de grote nationale en internationale fondsen voor onderzoeksfinanciering indien selectiecriteria worden gehanteerd die sturen in de richting van onderzoek dat wordt aangevraagd door grote onderzoeksconsortia en dat gericht is op innovatie en *high-risk*/*high-gain*, eerder dan op solide, beheersbare en betrouwbare wetenschap (Chhin, Taylor, & Wei, 2018; Falk-Krzesinski & Tobin, 2015; Koninklijke Nederlandse Akademie van Wetenschappen, 2018; Lilienfeld, 2017; Zwaan et al., 2018).

In de volgende paragrafen bespreken we de QRPs dan ook met dien verstande dat de oorzaken waarschijnlijk meervoudig en meerlagig zijn. Bovendien impliceert de (gedeeltelijke en gedeelde) verantwoordelijkheid van individuele onderzoekers en onderzoeksteams nog niet dat ook de remedie voor de repliceerbaarheidsproblemen bij de individuele onderzoekers en onderzoeksteams moet worden gezocht (zie Hoofdstuk 5).

3.3.1. Onbetrouwbare proefopzet, dataverzameling en data-analyse

Het eerste type van QRPs heeft te maken met de onbetrouwbaarheid bij de planning, uitvoering en analyse van de gegevens die zijn verzameld met empi-

risch onderzoek. Als het gaat over de onbetrouwbaarheid bij de planning van empirisch onderzoek, wordt een laag statistisch onderscheidingsvermogen bij de meerderheid van de studies in een bepaald onderzoeksdomein dikwijls als een oorzaak van repliceerbaarheidsproblemen vermeld (Bishop, 2019; Button et al., 2013; Ioannidis, 2005b; Open Science Collaboration, 2015). In een grootschalig overzicht van 12.065 schattingen van effectgroottes uit 200 meta-analyses en bijna 8000 publicaties wordt het lage onderscheidingsvermogen, samen met de hoge heterogeniteit van de effecten, in psychologisch onderzoek zelfs als een voldoende verklaring voor de huidige repliceerbaarheidsproblemen aangeduid (Stanley, Carter, & Doucouliagos, 2018).

De prominente rol die het statistisch onderscheidingsvermogen voor de verklaring van repliceerbaarheidsproblemen krijgt, kan vreemd lijken, omdat dit vermogen bij statistische toetsen wordt gedefinieerd als de kans om de nulhypothese te verwerpen, *gegeven dat de nulhypothese niet waar is* (het complement van de kans op een fout van de tweede soort; Neyman & Pearson, 1928, 1933; Cohen, 1962, 1969). Een laag statistisch onderscheidingsvermogen op zichzelf laat dus geen uitspraak toe over de kans om de nulhypothese te verwerpen, *gegeven dat de nulhypothese wel waar is* (de kans op een fout van de eerste soort). Het is namelijk juist die kans op een fout van de eerste soort die in het geding is bij repliceerbaarheidsproblemen. Gebrek aan repliceerbaarheid suggereert immers dat de oorspronkelijke studie een effect rapporteert (d.i. de nulhypothese verwerpt) dat niet bestaat (d.i. de nulhypothese is waar). Het is de verdienste van Ioannidis (2005b) en Button et al. (2013) te hebben benadrukt dat een laag onderscheidingsvermogen in interactie met andere QRPs, zoals hoge flexibiliteit bij de statistische analyse en *publicatiebias*, toch tot een verhoogd risico op de aanwezigheid van fouten van de eerste soort in de wetenschappelijke literatuur aanleiding geeft.

Deze koppeling van een laag onderscheidingsvermogen aan repliceerbaarheids-problemen is weinig verwonderlijk. Het onderscheidingsvermogen is gedefinieerd voor statistische hypothesetoetsen, maar binnen een algemener kader voor statistische inferentie verwijst een lager onderscheidingsvermogen naar lagere betrouwbaarheid, die bijvoorbeeld zichtbaar is in een grotere onnauwkeurigheid bij parameterschattingen (bredere betrouwbaarheidsintervallen) (Moore, McCabe, & Craig, 2017). Een lage betrouwbaarheid kan door een brede waaier van oorzaken ontstaan, bijvoorbeeld door een te klein aantal onafhankelijke observaties, onbetrouwbare meetinstrumenten, onzorgvuldige steekproeftrekking of de afwezigheid van (adequate) randomisering en *matching*. De implicatie is dat er grote schommelingen te verwachten zijn van de ene studie naar de andere, zelfs als het over perfecte directe replicaties zou gaan (De Boeck & Jeon, 2018; Loken & Gelman, 2017; Segers, 1989; Shadish et al., 2002). Grote schommelingen, in combinatie met het vertekend of selectief rapporteren van de 'positieve' bevindingen, zorgen voor niet-repliceerbare vals positieven in de wetenschappelijke literatuur

(Ioannidis, 2005b; Simmons et al., 2011). Maxwell waarschuwde in 2004 al voor deze gevaarlijke cocktail van een laag onderscheidingsvermogen, een hoge flexibiliteit bij de statistische analyse en andere methodologische problemen:

> 'Unless psychologists begin to incorporate methods for increasing the power of their studies, the published literature is likely to contain a mixture of apparent results buzzing with confusion. Increased reporting of effect sizes and confidence intervals will not by itself increase the consistency of the literature, although it may motivate more powerful studies by highlighting a major source of likely confusion. Not only do underpowered studies lead to a confusing literature but they also create a literature that contains biased estimates of effect sizes.' (p. 161)

Het blijft vreemd dat we anno 2020 nog steeds op het problematisch lage statistische onderscheidingsvermogen van empirisch onderzoek in de menswetenschappen moeten wijzen, een probleem dat al in 1962 door Cohen aan de kaak werd gesteld. In zijn baanbrekende overzichtsartikel kwam hij tot de conclusie dat het onderscheidingsvermogen in gepubliceerde gedragswetenschappelijke studies slechts 48% bedroeg bij een feitelijk effect van gemiddelde grootte. Dat is veel te laag, omdat het impliceert dat de kans op een fout van de tweede soort bijna tien keer zo hoog ligt als de nominaal vooropgestelde kans op een fout van de eerste soort. Het is moeilijk denkbaar dat onderzoekers de twee soorten fouten zo disproportioneel zouden wegen. Meer dan 25 jaar later bleek er bovendien weinig beterschap te zijn (Rossi, 1990; Sedlmeier & Gigerenzer, 1989) en nog eens een kwarteeuw later is er weliswaar een beetje vooruitgang in bepaalde domeinen en subdisciplines, maar blijft het geschatte gemiddelde onderscheidingsvermogen toch bedroevend laag (Dumas-Mallet, Button, Boraud, Gonon, & Munafò, 2017; Fraley & Vazire, 2014; Ioannidis et al., 2017; Stanley et al., 2018; Tressoldi, 2012; Tressoldi & Giofré, 2015; Vankov, Bowers, & Munafò, 2014).

Naast het lage statistische onderscheidingsvermogen van de gebruikte proefopzetten en de grote onnauwkeurigheid van de procedures die worden gebruikt bij de dataverzameling, is er ook nog een overvloed aan data-analytische technieken die tot vals positieve resultaten kunnen leiden. Dikwijls vermelde voorbeelden zijn: het verwijderen van uitschieters, het transformeren van variabelen en het meervoudig toetsen (Lindsay, 2015; Simmons et al., 2011; Wicherts et al., 2016).

1. Verwijderen van uitschieters

Het verwijderen van uitschieters is soms nodig als er fouten in de dataverzameling zijn geslopen of als er een vergissing bij de codering van de resultaten is opgetreden (Barnett & Lewis, 1994). Men kan een dergelijke uitschieter op het spoor komen door de frequentieverdeling van de resultaten te bekijken en dit te combineren met kwalitatieve informatie over de dataverzameling of de codering.

Als een bepaalde waarde ver van alle andere waarden verwijderd ligt (of soms helemaal onmogelijk is volgens de gehanteerde meetschaal) én er redenen zijn om aan te nemen dat het gaat over een dataverzamelings- of registratiefout, wordt meestal aangeraden om deze waarde te verwijderen en zo de gebruikelijke statistische procedures niet te vertekenen (Bollen & Jackman, 1990).

Het verwijderen van uitschieters kan echter ook een minder nobel doel dienen. Als er geen duidelijke afspraken werden gemaakt over het criterium om een waarde als 'ver van alle andere waarden verwijderd' te beschouwen en de betekenis van 'dataverzamelings- of registratiefout' wordt breed geïnterpreteerd, dan bestaat de verleiding om observaties die 'bizar' zijn of niet in lijn liggen met de verwachtingen van de onderzoeker onder het tapijt te vegen (Banks, Rogelberg, Woznyj, Landis, & Rupp, 2016; Matthes et al., 2015). Het spreekt voor zich dat de willekeur die op deze manier ontstaat, kan leiden tot vals positieven (Bakker & Wicherts, 2014b).

Om de negatieve effecten van het verwijderen van uitschieters te onderzoeken voerden Bakker en Wicherts (2014a) een literatuurstudie uit: ze vergeleken 92 artikels waarin uitschieters werden verwijderd met 61 artikels waarin geen verwijdering van uitschieters stond vermeld. Ze vonden geen statistisch significant verschil tussen de twee types van artikels in termen van *p*-waarden, steekproefgroottes en rapporteringsfouten. Dit suggereert dat het effect van het verwijderen van uitschieters op de aanwezigheid van vals positieven in de wetenschappelijke literatuur, en bij uitbreiding op het repliceerbaarheidsprobleem, eerder beperkt is. Bakker en Wicherts (2014a) stootten wel op een discrepantie tussen enerzijds de gerapporteerde vrijheidsgraden van de statistische toetsen en anderzijds de gerapporteerde steekproefgrootte in 41% van de artikels waarin geen verwijdering van uitschieters stond vermeld. Dit suggereert dan weer een hoge prevalentie van tekortkomingen in de rapportering van het verwijderen van uitschieters (of ontbrekende gegevens) en maakt de vergelijking van de twee types van artikels minder overtuigend. Het blijft dus niet uitgesloten dat de grote flexibiliteit bij het omgaan met uitschieters, in combinatie met andere QRPs, mee aan de basis van de huidige repliceerbaarheidsproblemen ligt.

2. Datatransformaties

Het verwijderen van uitschieters kan worden beschouwd als een extreme vorm van datatransformatie waarbij aan bepaalde waarden het gewicht 0 wordt gegeven. Om aan de assumpties van de gebruikelijke statistische procedures te voldoen worden soms ook veel algemenere datatransformaties voorgesteld, zoals logaritmische of vierkantsworteltransformaties (Bland & Altman, 1996a, 1996b; Moore et al., 2017). Er kunnen statistisch legitieme redenen zijn om datatransformaties uit te voeren, maar het grote aantal mogelijkheden om gegevens om te zetten en met elkaar te combineren (bv. in somscores of principale componentscores) zorgt ook wel voor een immens groot aantal *researcher degrees of freedom* (Wicherts et al., 2016).

Deze flexibiliteit bij de data-analyse kan tot 'datamassage' verworden als hierover niet op voorhand duidelijke criteria in een protocol werden vastgelegd (Nosek, Ebersole, DeHaven, & Mellor, 2018; Nosek et al., 2019). Als datatransformaties volledig optioneel en vrij zijn, kan elk effect statistisch significant worden gemaakt (Martin & Williams, 2017; Simmons et al., 2011).

3. Meervoudig toetsen

Meervoudig toetsen treedt op als er meerdere statistische toetsen op dezelfde dataset worden uitgevoerd (Benjamini & Hochberg, 1995; Miller, 1981; Westfall & Young, 1993). Meervoudig toetsen is op zich geen probleem, maar het wordt er wel een als er geen rekening wordt gehouden met de veelheid aan mogelijkheden die het met zich meebrengt en met de verhoging van de kans dat er ten minste één statistisch significant effect wordt gevonden, zelfs als alle nulhypothesen waar zijn. Meervoudig toetsen wordt dus problematisch als het uitmondt in opportunistisch toetsen en het selectief rapporteren van die toetsresultaten die aan de verwachtingen van de onderzoeker voldoen (Maxwell, 2004; Simmons et al., 2011; Wicherts et al., 2016).

Een bijzondere vorm van meervoudig toetsen ontstaat als de dataverzameling over een langere periode verloopt en het stoppen van het onderzoek afhankelijk wordt gemaakt van het bereiken van statistische significantie, ook wel *optional stopping* genoemd (Fisher, 2014; Schott, Rhemtulla, & Byers-Heinlein, 2019; Wagenmakers, 2007). Een onderzoeker die deze strategie gebruikt zal een aantal deelnemers rekruteren, de gegevens analyseren en alleen overgaan tot het afsluiten van de studie, redactie en publicatie als het gewenste resultaat is bereikt. Zolang dat niet het geval is, zal de onderzoeker doorgaan met het rekruteren van deelnemers en het analyseren van gegevens tot het resultaat er is (bv. als er 'statistische significantie' wordt bereikt) of tot de tijd, de energie en het geld op zijn. De tussentijdse toetsen die worden uitgevoerd zijn vergelijkbaar met de meervoudige toetsen op dezelfde dataset en zorgen op dezelfde manier voor een inflatie van het risico op vals positieven (Lindsay, 2015; Sanborn & Hills, 2014; Simmons et al., 2011).

Bij de statistische analyse kan men wel rekening houden met het gebruik van deze *optional stopping*-strategie, zowel binnen het kader van de frequentistische statistiek (Armitage, McPherson, & Rowe, 1969; Schott et al., 2019; Wald, 1947) als van de Bayesiaanse statistiek (Schonbrodt, Wagenmakers, Zehetleitner, & Perugini, 2017; Wagenmakers, 2007). Het probleem ontstaat als een onderzoeker zijn flexibele manier van werken met het aantal deelnemers 'verpakt' als een onderzoek met een op voorhand vastliggend steekproefplan met een vast aantal deelnemers op basis van een a-priorianalyse van het onderscheidingsvermogen. Als onderzoekers de inschatting van het risico op vals positieven niet aan de gebruikte strategie van dataverzameling aanpassen, misleiden zij zichzelf. En als

zij op die manier de resultaten in de wetenschappelijke literatuur rapporteren, wordt samen met hen alle lezers een rad voor de ogen gedraaid (Lindsay, 2015; Sanborn & Hills, 2014; Simmons et al., 2011). Merk ten slotte op dat we hier de discussie over meervoudig toetsen en optioneel stoppen, met het oog op de eenvoud, in termen van statistische toetsen hebben gevoerd, maar dat een analoge discussie mogelijk is voor simultane betrouwbaarheidsintervallen bij het schatten van effectgroottes (Belia, Fidler, Williams, & Cumming, 2005; Cumming & Maillardet, 2006; Maxwell, 2004; Morey, Hoekstra, Rouder, Lee, & Wagenmakers, 2016).

3.3.2. Vertekende rapportering

Het zorgvuldig rapporteren van empirisch onderzoek is cruciaal voor de repliceerbaarheid. Hoewel er een hoge onbetrouwbaarheid in de proefopzet, de dataverzameling en de data-analyse kan sluipen, kan er veel worden opgelost door open en transparant over alle onderzoeksstappen te rapporteren. Een vertekende rapportering geeft lezers, het brede publiek en professionals die de bevindingen in de praktijk willen omzetten, een verkeerde indruk van de onzekerheid die nog met de bevindingen gepaard gaat. Meer in het bijzonder zet een vertekende rapportering andere onderzoekers op het verkeerde been wat de repliceerbaarheid van de bevindingen betreft (Aczel et al., 2019; Nosek et al., 2015; Simmons et al., 2011).

Een van de meest funeste manieren om met een vertekende rapportering de repliceerbaarheid te ondergraven is het voorstellen van exploratief onderzoek als toetsend onderzoek (Wagenmakers et al., 2012). In dat geval wordt de indruk gewekt dat er een rechtlijnige afleiding van theorie naar hypothese en van hypothese naar predicties werd gemaakt, en dat het onderzoek dus door een hoge validiteit en betrouwbaarheid wordt gekenmerkt. Als de hypothese echter tot stand is gekomen door eerst de data te analyseren en vervolgens op zoek te gaan naar een ad-hoctheorie of een hypothese die de data verklaart (terwijl de omgekeerde volgorde wordt gerapporteerd), wordt er geen rekening gehouden met alle mogelijkheden waarop data tot stand kunnen komen en met het veelvoud aan mogelijkheden om hypothesen te vinden die perfect bij empirische data in menswetenschappelijk onderzoek passen (Rubin, 2017; Wagenmakers et al., 2012).

Kerr (1998) nam deze volgens hem courante manier van vertekening in de rapportering van empirisch menswetenschappelijk onderzoek op de korrel. Hij introduceerde het ondertussen ingeburgerde acroniem *HARKing*, dat staat voor *Hypothesizing After the Results are Known*. HARKen komt in verschillende vormen en gradaties voor, bijvoorbeeld door de resultaten te gebruiken om post-hoc-hypothesen op te stellen die als a-priori-hypothesen worden voorgesteld, door hypothesen af te leiden uit een post-hoc-literatuurstudie die als a-prior-

hypothesen worden vooropgesteld, of door over a-priori-hypothesen die in strijd zijn met de geobserveerde resultaten in het onderzoeksrapport te zwijgen (Bishop, 2019; Rubin, 2017).

Het is belangrijk op te merken dat HARKen op zich geen probleem vormt. Speculeren op basis van de getoetste hypothesen in confrontatie met de verzamelde data maakt zelfs meestal deel uit van het onderdeel 'Suggesties voor verder onderzoek' in de discussiesectie van een wetenschappelijk artikel. Het problematische aspect is gelegen in het herschrijven van de introductiesectie en de probleemstelling van een wetenschappelijk artikel op basis van de geobserveerde resultaten. Anders gezegd: HARKen maakt deel uit van de empirische cyclus, maar de door data geïnspireerde hypothesen zouden met een nieuwe onafhankelijke dataverzameling moeten worden getoetst (Hollenbeck, & Wright, 2017; Murphy & Aguinis, 2019).

Problematische vormen van HARKen blijken een hoge incidentie te hebben. Afhankelijk van de manier waarop de vraag wordt gesteld en van het rechtstreeks bevragen of het onrechtstreeks afleiden van sporen van HARKen, wordt het bij 30 tot 70% van de onderzoekers in de empirische menswetenschappen aangetroffen (Murphy & Aguinis, 2019). Deze hoge incidentie is niet verwonderlijk omdat HARKen deel lijkt uit te maken van de cultuur waarin onderzoekers worden opgeleid en de manier waarop hen wordt aangeleerd om publiceerbare wetenschappelijke artikels te schrijven. In dat opzicht is hoofdstuk 10, *Writing the Empirical Journal Article*, van *The compleat academic: A practical guide for the beginning social scientist* (Darley, Zanna, & Roediger, 2004), uitgegeven door de *American Psychological Association*, ontluisterend. Het werd geschreven door Daryl Bem (2004) en steekt onder de hoofding *Which Article Should You Write?* van wal met het volgende niet mis te verstane advies:

> 'There are two possible articles you can write: (a) the article you planned to write when you designed your study or (b) the article that makes the most sense now that you have seen the results. They are rarely the same, and the correct answer is (b).' (p. 186)

Verder in het hoofdstuk staat expliciet: '(…) the data may be strong enough to justify recentering your article around the new findings and subordinating or even ignoring your original hypotheses.' (p. 187). En ook: 'If your results suggest a compelling framework for their presentation, adopt it and make the most instructive finding your centerpiece.' (p. 188)

Het eindrapport wordt dus een goed verhaal, maar tegelijk ook een opgesmukte en verbloemde weergave van het wetenschappelijk proces. Op de achterflap van het boek lezen we nog: 'This book provides invaluable guidance that will help new academics plan, play, and ultimately win the academic career game.'[6] Het

[6] Zie ook https://www.apa.org/pubs/books/4316014.

is dubbel ironisch dat de auteur van dit hoofdstuk een aantal jaar later in een vlaggenschiptijdschrift van de *American Psychological Association* negen niet-repliceerbare experimenten publiceerde die aantoonden dat mensen beïnvloed kunnen worden door een onvoorspelbare gebeurtenis in de toekomst (Bem, 2011).

3.3.3. Selectieve rapportering

Selectieve rapportering kan worden beschouwd als een extreme vorm van vertekende rapportering. Bij selectieve rapportering wordt er niet louter vervormd, opgesmukt of verbloemd, maar wordt belangrijke informatie uit de rapportering en de gepubliceerde literatuur weggelaten. We bespreken achtereenvolgens *publicatiebias*, *p-hacking*, en *outcome switching*.

1. Publicatiebias

Selectieve rapportering op het niveau van de individuele publicatie is een soort van vertekende rapportering, bijvoorbeeld als HARKen de vorm krijgt van het verzwijgen van a-priori-hypothesen die in strijd zijn met de geobserveerde resultaten (Kerr, 1998; Rubin, 2017). Ook selectieve rapportering op het niveau van een heel onderzoeksdomein komt voor en is zelfs een wezenlijk deel van het publicatieproces. In dit proces zitten namelijk noodzakelijk een aantal filters die studies met methodologische tekortkomingen of rapporteringsfouten uitsluiten of tijdelijk uitsluiten, vergezeld van suggesties voor verbetering. Selectieve rapportering op het niveau van een onderzoeksdomein wordt problematisch voor de repliceerbaarheid als bij *reviewers*, *editors* of de onderzoekers zelf motieven voor niet-publicatie beginnen mee te spelen die van inhoudelijke of ideologische aard zijn of met de geobserveerde data zelf te maken hebben (Ellemers, 2013; Hopf et al., 2019; Lindsay, 2015).

Publicatiebias is een vertekening die voor een heel onderzoeksdomein ontstaat door een selectieve rapportering die inderdaad met de geobserveerde data zelf te maken heeft. In de klassieke betekenis van *publicatiebias* gaat het over de selectieve publicatie van studies die statistisch significante resultaten opleveren. Deze praktijk is ingegeven door de misvatting dat uit studies met niet-statistisch significante resultaten niets te leren valt of dat in die studies bepaalde onvolkomenheden aanwezig zijn, waardoor ze niet voor publicatie vatbaar zijn (Bishop, 2019; Franco, Malhotra, & Simonovits, 2014; Greenwald, 1975; Rosenthal, 1979).

Statistici waarschuwden in het midden van de vorige eeuw reeds voor de nefaste gevolgen van deze manier om publicaties te selecteren:

> ‘There is some evidence that in fields where statistical tests of significance are commonly used, research which yields nonsignificant results is not

> published. Such research being unknown to other investigators may be repeated independently until eventually by chance a significant result occurs – an "error of the first kind" – and is published. Significant results published in these fields are seldom verified by independent replication. The possibility thus arises that the literature of such a field consists in substantial part of false conclusions resulting from errors of the first kind in statistical tests of significance.' (Sterling, 1959, p. 30)

Ook psychologen erkennen dit probleem sinds lang:

> 'For any given research area, one cannot tell how many studies have been conducted but never reported. The extreme view of the "file drawer problem" is that journals are filled with the 5% of the studies that show Type I errors, while the file drawers are filled with the 95% of the studies that show nonsignificant results.' (Rosenthal, 1979, p. 638)

Beide citaten lijken een voorafspiegeling van *Why Most Published Research Results Are False* (Ioannidis, 2005b). *Publicatiebias* kan rechtstreeks met repliceerbaarheidsproblemen in verband worden gebracht, omdat het meervoudig onderzoeken van een effect tot een uitzonderlijke toevalstreffer kan leiden, zelfs als er geen enkel effect bestaat. Deze toevalstreffer wordt gepubliceerd, maar is onmogelijk te repliceren tenzij er weer een veelvoud aan replicatiestudies uitgeprobeerd wordt.

Er is overigens een overvloed aan empirische evidentie dat *publicatiebias* wel degelijk in verschillende onderzoeksdomeinen aanwezig is (Francis, 2012, 2014; Franco et al., 2014; Sterling, Rosenbaum, & Weinkam, 1995; Rothstein, Sutton, & Borenstein, 2005). Bovendien zijn er aanwijzingen dat auteurs meer en meer anticiperen op publiceerbaarheid en bijgevolg studies met niet-statistisch significante bevindingen zelfs niet uitschrijven (of ze beperken tot een masterproef of een intern rapport, en dus niet insturen bij een wetenschappelijk tijdschrift; zie Cooper, DeNeve, & Charlton, 1997; Dickersin, Chan, Chalmers, Sacks, & Smith Jr., 1987; Dickersin, Min, & Meinert, 1992; Franco et al., 2014). De selectieve rapportering wordt op deze manier een vorm van zelfselectie of zelfcensuur.

Merk op dat *publicatiebias* in de klassieke opvatting aan statistische significantie wordt gekoppeld (Greenwald, 1975; Rosenthal, 1979; Sterling, 1959). Hierdoor ontstaat de indruk dat de significantietoets (of het verkeerd gebruik ervan) voor de *publicatiebias* verantwoordelijk moet worden gesteld. In een bredere definitie van *publicatiebias* gaat het echter over elk systematisch verschil tussen gepubliceerde en niet-gepubliceerde studies. Meestal hebben studies die een groter dan gemiddeld effect vinden meer kans om gepubliceerd te worden (Borenstein, Hedges, Higgins, & Rothstein, 2009). Dit zorgt voor een positieve vertekening van gepubliceerde effecten en biedt meteen ook een afdoende verklaring voor

het kleinere effect in replicatiestudies dan in de oorspronkelijke studies (Camerer et al., 2016, 2018; Ebersole et al., 2016; Klein, 2014, 2018, 2019; Open Science Collaboration, 2015).

2. P-hacking

P-hacking is een term die verwijst naar het veelvuldig uitvoeren van statistische significantietoetsen binnen één studie, bijvoorbeeld met verschillende mogelijke uitkomstmaten, waarbij alleen de toetsen die de gewenste lage *p*-waarde opleveren (d.i. die statistisch significant zijn volgens de conventie van het veld) worden gerapporteerd (Bishop, 2019; Nelson et al., 2018; Simmons et al., 2011; Wicherts et al., 2016). Ook hierbij moeten we opmerken dat er weliswaar verwezen wordt naar de *p*-waarde van klassieke significantietoetsen, maar dat *p-hacking* in een brede betekenis ook toepasbaar is bij het gebruik van andere statistische technieken (bv. betrouwbaarheidsintervallen en Bayesiaanse analyses). Die bredere betekenis wordt gevat door de synoniemen *data dredging* en *cherry-picking* van onderzoeksresultaten (Marín-Franch, 2018).

P-hacking is verwant aan het meervoudig toetsingsprobleem uit 3.3.1, maar daar ging het over het miskennen van een nood tot statistische correctie voor het meervoudig toetsen. Bij *p-hacking* gaat het in essentie over een probleem van selectieve rapportering. Phillips (2004) noemde het dan ook terecht 'publication bias in situ'.

Er zijn aanwijzingen dat *p-hacking* een wijdverspreide onderzoeksstrategie is (Head, Holman, Lanfear, Kahn, & Jennions, 2015; John et al., 2012; Nelson et al., 2018; Simmons et al., 2011). *P-hacking* verklaart de ogenschijnlijke paradox dat enerzijds studies in de empirische menswetenschappen een te klein onderscheidingsvermogen hebben en dat anderzijds tijdschriften bol staan van de statistisch significante resultaten. Voor Simmons et al. (2011) en Nelson et al. (2018) vormt *p-hacking* de belangrijkste methodologische verklaring voor repliceerbaarheidsproblemen.

Een schoolvoorbeeld van *p-hacking* vinden we in het werk van voedingspsycholoog Brian Wansink (Steijaert, 2017, 2018). Wansink was een vermaard expert en mediafiguur op het gebied van eetgedrag en het bestrijden van zwaarlijvigheid door het veranderen van omgevingsfactoren die ons eetgedrag beïnvloeden (bv. door uit kleinere borden te eten). Hij schreef de bestseller *Mindless eating: Why we eat more than we think* (Wansink, 2006), was hoogleraar aan Cornell University en stond van 2007 tot 2009 aan het hoofd van het Center for Nutrition Policy and Promotion van het United States Department of Agriculture, de laatste twee jaar onder president George W. Bush (Lawn, 2011).

In 2016 startte hij een blog met als allereerste artikel *The Grad Student Who Never Said "No"* (21 november 2016).[7] Dat artikel gaat over een Turkse doctoraatsstudente die voor zes maanden als vrijwillig medewerker in Wansinks lab komt werken. Het artikel wil een boodschap meegeven over de deugdzaamheid van hard en onbetaald werken en het grijpen van de kansen die zich in het leven voordoen. Tegelijk onthult het de *p-hacking*-strategie die in het lab wordt gehanteerd:

> 'When she arrived, I gave her a data set of a self-funded, failed study which had null results (it was a one month study in an all-you-can-eat Italian restaurant buffet where we had charged some people ½ as much as others). I said, "This cost us a lot of time and our own money to collect. There's got to be something here we can salvage because it's a cool (rich & unique) data set." I had three ideas for potential Plan B, C, & D directions (since Plan A had failed). I told her what the analyses should be and what the tables should look like. I then asked her if she wanted to do them.
> Every day she came back with puzzling new results, and every day we would scratch our heads, ask "Why," and come up with another way to reanalyze the data with yet another set of plausible hypotheses. Eventually we started discovering solutions that held up regardless of how we pressure-tested them. I outlined the first paper, and she wrote it up, and every day for a month I told her how to rewrite it and she did. This happened with a second paper, and then a third paper (which was one that was based on her own discovery while digging through the data).'

Op het artikel kwamen heel veel reacties, die Wansink aanvankelijk als heel constructief interpreteerde. Zijn respons maakte de critici nog achterdochtiger:

> 'P-hacking and MTurk-iterating isn't helpful to science, and it's one of the reasons our lab seldom cites on-line studies. However, P-hacking shouldn't be confused with deep data dives – with figuring out why our results don't look as perfect as we want.
> With field studies, hypotheses usually don't "come out" on the first data run. But instead of dropping the study, a person contributes more to science by figuring out when the hypo worked and when it didn't. This is Plan B. Perhaps your hypo worked during lunches but not dinners, or with small groups but not large groups. You don't change your hypothesis, but you figure out where it worked and where it didn't. Cool data contains cool discoveries. If a pilot study didn't precede the field study, a lab study can follow -- either we do it or someone else does.'

De 'deep data dives' overtuigden niet. Wat Wansink in het artikel beschreef, is perfect legitiem als exploratieve strategie, maar het probleem is dat in zijn

[7] De originele blog is intussen verwijderd, maar werd gearchiveerd als https://web.archive.org/web/20170312041524/http://www.brianwansink.com/phd-advice/the-grad-student-who-never-said-no.

artikels alleen melding wordt gemaakt van de analyses die de gewenste resultaten opleveren, niet van de andere.

Dit voorbeeld heeft overigens geen goede afloop. De betreffende artikels werden nagekeken en bleken bol te staan van andere rapporteringsfouten en inconsistenties (Van der Zee, Anaya, & Brown, 2017). Ook vroegere publicaties van Wansink werd uitgevlooid. De ene na de andere bleek fundamentele fouten te bevatten en moest worden teruggetrokken (Munafò, Hollands, & Marteau, 2018). Een interne onderzoekscommissie van Cornell University onderzocht vervolgens het werk van Wansink en kwam met een vernietigend verslag waarin de academicus schuldig werd bevonden aan 'academic misconduct in his research and scholarship, including misreporting of research data, problematic statistical techniques, failure to properly document and preserve research results, and inappropriate authorship' (Lee, 2018). Hij werd op non-actief gesteld. Uiteindelijk hield Wansink de eer aan zichzelf door op 30 juni 2019 ontslag te nemen (Lee, 2018; Munafò et al., 2018; Steijaert, 2018).

3. Outcome switching

Outcome switching is een soort van selectieve rapportering waarbij de oorspronkelijke afhankelijke variabele wordt ingeruild voor een andere, waarvoor de resultaten gunstiger zijn (Chan, Hróbjartsson, Haahr, Gøtzsche, & Altman, 2004; Claesen, Gomes, Tuerlinckx, & Vanpaemel, 2019; Goldacre, 2016). Dit kan het resultaat zijn van *p-hacking*, maar dat hoeft niet.

Outcome switching kan men op het spoor komen door onderzoeksprotocollen te vergelijken met corresponderende publicaties. Het gebruik van gepreregistreerde onderzoeksprotocollen is bij klinische studies al langer ingeburgerd. Bij de vergelijking tussen die protocollen en de uiteindelijke publicaties blijkt er een aanzienlijk aantal afwijkingen te zijn, onder meer op het vlak van de primaire uitkomstmaat, zonder dat hierover transparant wordt gerapporteerd (Altman, Moher, & Schulz, 2017; Claesen et al., 2019; Jones, Keil, Holland, Caughey, & Platts-Mills, 2015). Als het gaat over het verwisselen van de primaire en secundaire uitkomstmaat, kan *outcome switching* ook als een soort vertekende rapportering worden beschouwd – analoog aan *hypothesis switching* bij HARKing – maar evengoed gaat het over het toevoegen van afhankelijke variabelen in de publicatie die nergens in het protocol vermeld staan, of over het verwijderen van afhankelijke variabelen in de publicatie die wél in het protocol vermeld staan.

Zoals andere vormen van vertekende of selectieve rapportering verhoogt *outcome switching* het risico op vals positieven (Wicherts, 2017). In meta-analyses heeft het als gevolg dat er een vertekening is in de richting van de effecten die onderzoekers graag zouden zien optreden (Altman et al., 2017; Jones et al., 2015; Hengartner, 2018).

4. Implicaties

Repliceerbaarheidsproblemen, al dan niet in combinatie met bovenstaande QRPs, hebben implicaties voor de reputatie van individuele wetenschappers, maar ook van de empirische menswetenschappen als geheel, en bij uitbreiding van de hele academische en wetenschappelijke wereld. Wetenschappers zijn echter niet bij de pakken blijven zitten en hebben de repliceerbaarheidsproblemen aangegrepen om hervormingen voor te stellen en in 'het wetenschapsbedrijf' doortastende veranderingen door te voeren. Op deze verschillende implicaties gaan we in de volgende secties kort in.

4.1. Implicaties voor individuele wetenschappers

Repliceerbaarheidsproblemen kunnen de reputatie van individuele wetenschappers blijvend bepalen. Zoals we in paragraaf 3.3.3 illustreerden met het onderzoek van Brian Wansink, kunnen zelfs QRPs die worden ontdekt in een beperkt aantal publicaties, aanleiding geven tot een doorlichting van een hele publicatielijst, met als mogelijk gevolg een hele lawine van teruggetrokken publicaties en zelfs ontslag.

Bij Daryl Bem (zie 3.1 en 3.3.2), ook van Cornell University, is het anders gelopen. Bem was 73 en al met emeritaat toen het beruchte *Feeling the future* (Bem, 2011) verscheen. Bovendien is hij dit onderzoek blijven verdedigen tegen methodologische en inhoudelijke kritiek (Bem, Tressoldi, Rabeyron, & Duggan, 2015/2016). Recente uitspraken van Bem maken zijn verdediging echter weinig overtuigend:

> 'I'm all for rigor, but I prefer other people do it. I see its importance – it's fun for some people – but I don't have the patience for it. If you looked at all my past experiments, they were always rhetorical devices. I gathered data to show how my point would be made. I used data as a point of persuasion, and I never really worried about, "Will this replicate or will this not?"' (Daryl Bem, in Engber, 2017)

Voor *tenure-track*-docenten kunnen repliceerbaarheidsproblemen en aanhoudende methodologische kritiek echter fataal zijn. Dit overkwam Amy Cuddy, coauteur van een artikel over *power posing* dat veel stof deed opwaaien (Carney, Cuddy, & Yap, 2010). *Power posing* verwijst naar het aannemen van een lichaamshouding die aan gevoelens van psychologische kracht en assertiviteit worden gekoppeld (bv. de handen in de zij en de benen licht uit elkaar). Carney et al. (2010) hadden 42 deelnemers op een zuiver toevallige wijze in twee groepen verdeeld: de ene groep moest twee keer een minuut lang een krachtige lichaamshouding aannemen, de andere een verkrampte lichaamshouding. De onderzoekers vonden in de eerste groep, in vergelijking met de tweede, niet alleen een statistisch

significant hoger gemiddelde in de zelfrapportering van psychologische kracht, maar ook statistisch significant hogere niveaus van testosteron (geïnterpreteerd als een 'dominantiehormoon'), lagere niveaus van cortisol (geïnterpreteerd als een 'stresshormoon') en een meer rationeel risicogedrag (in een financiële goktaak).

In 2012 pakte Amy Cuddy met dit resultaat uit in een TED Talk[8]: *Your body language may shape who you are*. Die werd ondertussen miljoenen keer bekeken. Ze schreef ook de bestseller *Presence: Bringing your boldest self to your biggest challenges* (Cuddy, 2015), die ondertussen in 34 talen werd vertaald, en ze werd een graag geziene gast in talkshows. Haar '*life hack*', zoals ze het zelf noemt, is dan ook spectaculair: het volstaat om twee minuten lang een krachtige lichaamshouding aan te nemen en er zijn al meteen effecten voor het zelfvertrouwen, de hormonenproductie en het gedrag. Als er dankzij het verhoogde zelfvertrouwen een positieve prestatie volgt, kan er bij herhaling een grote stijging van het zelfwaardegevoel worden verwacht.

Het was te mooi om waar te zijn. In een replicatiestudie van Ranehill et al. (2015) met 200 deelnemers werd alleen het effect op de zelfrapportering teruggevonden. De effecten op het testosteronniveau, het cortisolniveau en de gedragstaken konden niet worden gerepliceerd. Bovendien werd Cuddy na deze mislukte replicatiepoging verrast door een publiek statement van haar coauteur Dana Carney, die op de website van de University of California, Berkeley postte: 'I do not believe that "power pose effects" are real... the evidence against the existence of power poses is undeniable.' Cuddy bleef de effecten verdedigen, maar ook Simmons en Simonsohn (2017), de auteurs van de *False-positive psychology*-paper uit 2011 (zie sectie 3.1), kwamen op basis van een statistische analyse van 33 experimenten tot deze conclusie:

> 'Taken together, the results from Ranehill et al.'s (2015) replication and from our *p*-curve analysis suggest that the behavioral and physiological effects of expansive versus contractive postures ought to be treated as hypotheses currently lacking in empirical support. Although more highly powered future research may find replicable evidence for those benefits (or unexpected detriments), the existing evidence is too weak to justify a search for moderators or to advocate for people to engage in power posing to better their lives.' (pp. 690–691)

Cuddy verliet haar *tenure-track*-positie aan de Harvard Business School in de lente van 2017 (Dominus, 2017).[9]

---

[8] https://www.youtube.com/watch?v=Ks-_Mh1QhMc; volgens https://books.google.be/books/about/Presence.html?id=RL8srgEACAAJ&redir_esc=y is het de tweede meest bekeken presentatie in de geschiedenis van de TED Talks.

[9] In de marge hiervan was er veel te doen over de negatieve effecten van 'methodologisch terrorisme' en *public shaming and blaming* op sociale media (Fiske, 2016). We willen de oproep

4.2. De geloofwaardigheid van de wetenschap

Repliceerbaarheidsproblemen zetten dus, behalve de reputatie van individuele wetenschappers, ook de geloofwaardigheid van de empirische menswetenschappen op het spel (Ferguson, 2015; Pashler & Wagenmakers, 2012). Bovendien zijn de repliceerbaarheidsproblemen koren op de molen van antiwetenschappelijke groeperingen die het vertrouwen in volledige wetenschappelijke disciplines, wetenschappelijke theorieën en onderzoeksgebaseerde praktijken onderuit willen halen. Denk aan het ter discussie stellen van de klimaatwetenschap, de plaats van de evolutietheorie in het biologieonderwijs en het verdacht maken van vaccinaties (Jamieson, 2018; Reygel, 2019; Schulson, 2018). Ten slotte zijn er ook aanwijzingen dat een beschadigd vertrouwen in de wetenschap niet eenvoudig door communicatie en beloftes over verhoogde methodologische standaarden te herstellen is (Wingen, Berkessel, & Englich, 2019). Het zal dus zaak zijn het bestaande vertrouwen in de wetenschap in stand te houden en dat ook blijvend waard te zijn (zie ook All European Academies, 2018c, 2019a, 2019b).

Het vertrouwen van het brede publiek in de wetenschap lijkt overigens nog altijd bijzonder groot. In de Verenigde Staten is het de afgelopen vijftig jaar stabiel gebleven, terwijl andere instituten, zoals het Congres, grote bedrijven en de pers, er aanzienlijk op achteruit zijn gegaan (National Science Foundation, 2018). Als baken van vertrouwen wordt wetenschap in de Verenigde Staten sinds de terreuraanslagen van 9/11 alleen voorafgegaan door het leger (National Science Foundation, 2018). Bij de meest recente enquête voor de *Science & Engineering Indicators* antwoordde 88% van de respondenten dat ze 'in sterke mate akkoord' of 'akkoord' gaan met de uitspraak dat '[m]ost scientists want to work on things that will make life better for the average person'. En 89% ging 'in sterke mate akkoord' of 'akkoord' met de uitspraak dat '[s]cientific researchers are dedicated people who work for the good of humanity' (National Science Foundation, 2018). Uit de jaarlijkse Ipsos-enquête blijkt dat wetenschappers in de rangschikking van de meest geloofwaardige beroepscategorieën wereldwijd nog altijd op de eerste plaats staan (60%), gevolgd door artsen (56%), leerkrachten (52%) en militairen (43%) (Skinner, 2019). In België bestaat de top vier uit artsen (64%), wetenschappers (56%), leerkrachten (51%) en politieagenten (44%) (Ipsos, 2019).

---

van Fiske om wetenschappelijke kritiek op een respectvolle, constructieve en *evidence-based* manier te formuleren dan ook onderschrijven. Cuddy zelf heeft trouwens buiten de academische wereld een succesvolle loopbaan als consultant en schrijfster uitgebouwd en ze blijft zich op een indrukwekkende en wetenschappelijke manier verweren (Cuddy, Schultz, & Fosse, 2018; Elsesser, 2018). Credé (2019) beargumenteert echter terecht dat haar verweer niet overtuigt en dat zelfs de zogenaamde power pose-effecten op zelfrapporteringen vooral toe te schrijven zijn aan het negatieve effect van de groep die de verkrampte lichaamshouding moest aannemen.

Scheufele (2014) wees erop dat ongefundeerde aandacht voor nieuwe en spectaculaire bevindingen het vertrouwen in de wetenschap aan het wankelen kan brengen, zeker als er hoge verwachtingen worden gewekt die vervolgens niet worden ingelost, of als initieel beloftevolle resultaten niet kunnen worden gerepliceerd. In een Amerikaanse opiniepeiling gaf 74% van de respondenten aan dat ze het een probleem vinden dat '[s]cience researchers overstate the implications of their research' (27% als een groot probleem en 47% als een klein probleem) (Funk et al., 2017). Met andere woorden: 'Science may run the risk of undermining its position in society in the long term if it does not navigate this area of public communication carefully and responsibly' (Scheufele & Krause, 2019, p. 6).

4.3. Repliceerbaarheidsproblemen als impuls voor verandering

De repliceerbaarheidsproblemen hebben negatieve gevolgen gehad – individueel (zie sectie 4.1) en collectief (zie sectie 4.2) – en ze hebben wetenschappers aangezet tot het zoeken naar verklaringen voor de niet-repliceerbaarheid (zie sectie 3.3). Maar ze hebben hen er ook toe aangezet om de gangbare onderzoekspraktijken en -culturen in onderzoeksinstituten en labo's ter discussie te stellen en met remedies voor de dag te komen (Rodgers & Shrout, 2018).

Als tegengewicht voor de negatieve of misleidende connotaties bij de term 'repliceerbaarheidscrisis' (zie paragraaf 2.2.1.) worden daarbij de grote woorden alweer niet geschuwd. Er is sprake van een 'renaissance' (Nelson et al., 2018) en een '*credibility revolution*' (Vazire, 2018). Met een verwijzing naar de voorgestelde remedies wordt er gesproken van een '*preregistration revolution*' (Nosek et al., 2018) en een '*cooperative revolution*' (Chartier et al., 2018). Preregistratie is het op voorhand in een protocol vastleggen van de onderzoeksplanning, -uitvoering en -analyse om QRPs te vermijden (zie 3.3.1), zoals bij klinische studies al langer het geval is. Coöperatie betreft diverse vormen van onderzoekssamenwerking om fouten, misleiding en vertekening te vermijden (Uhlmann et al., 2019; Wicherts, 2011). Om het hoofd te bieden aan de repliceerbaarheidsproblemen binnen de psychologie en om de remedies te onderzoeken en te bediscussiëren werd in 2016 *The Society for the Improvement of Psychological Science* en het tijdschrift *Collabra: Psychology* opgericht, op zich al een mooi staaltje van coöperatie.[10]

Er wordt wel eens beweerd dat zelfcorrectie een wezenlijk kenmerk van wetenschap is (Hilgard & Jamieson, 2017; Ioannidis, 2012; Merton, 1973). Dat is terecht en interessant, maar die zelfcorrectie werkt niet als een automaat die 'de wetenschap' op zichzelf uitvoert; wetenschap wordt door wetenschappers bedreven en het zijn wetenschappers die de correcties moeten doorvoeren (Oreskes, 2019). Aan

[10] https://improvingpsych.org/

het eind van dit Standpunt formuleren we daarom enkele aanbevelingen om de repliceerbaarheidsproblemen een daadwerkelijke impuls voor verandering te laten zijn.

5. Aanbevelingen

Er is al veel inkt gevloeid over de manier waarop we met de repliceerbaarheidsproblemen in de empirische menswetenschappen moeten omgaan (De Boeck & Jeon, 2018; Ioannidis, Munafò, Fusar-Poli, Nosek, & David, 2014; Koninklijke Nederlandse Akademie van Wetenschappen, 2018; LeBel et al., 2018; Munafò et al., 2017; National Academies of Sciences, Engineering, and Medicine, 2019; Nosek & Bar-Anan, 2012; Nosek, Spies, & Motyl, 2012; Shrout & Rodgers, 2018; The Academy of Medical Sciences, 2015; Uhlmann et al., 2019). Opvallend is de lange lijst en de grote verscheidenheid aan voorgestelde oplossingen. Bovendien lijkt de redenering soms te zijn: 'Er is een repliceerbaarheidsprobleem en de meeste wetenschappers gebruiken X. Laten we daarom stoppen met X en in plaats daarvan Y gebruiken (en dan is het repliceerbaarheidsprobleem opgelost)', waarbij Y het stokpaardje is waarvoor de hervormers promotie voeren.

Op basis van het overzicht en de analyse in de voorgaande secties kunnen we de grote verscheidenheid aan voorgestelde oplossingen tot drie algemene aanbevelingen terugbrengen. We beperken ons tot aanbevelingen die repliceerbaarheid als rechtstreeks doel hebben:

1. Ondersteun replicatieonderzoek.
2. Ondersteun open wetenschappelijk onderzoek.
3. Ondersteun zorgvuldig wetenschappelijk onderzoek.

Twee flankerende suggesties zouden we hieraan willen toevoegen:

4. Ondersteun methodologieonderwijs.
5. Ondersteun metaonderzoek.

In wat volgt becommentariëren we de vijf aanbevelingen en geven we aan hoe elke aanbeveling rechtstreeks op de repliceerbaarheidsproblemen inspeelt. Op de tweede aanbeveling – 'Ondersteun open wetenschappelijk onderzoek' – gaan we iets dieper in omdat in dat verband momenteel de meeste initiatieven worden genomen en ook het grootste aantal controverses bestaan.

Bij de aanbevelingen en flankerende suggesties gaan we niet in op de manier waarop de actoren, zoals de universiteiten, de onderzoeksfinanciers (o.m. FWO-Vlaanderen, de Vlaamse, Belgische en Europese overheden) en de tijdschriftredacties, hierbij concreet moeten worden betrokken. Het spreekt echter voor zich dat lippendienst of een schouderklopje niet volstaan om de 'Ondersteun…' in elke aanbeveling in de praktijk te brengen. Eén voorbeeld: met een gericht en expliciet promotiebeleid beschikken de universiteiten over een krachtig instrument om de beloningsstructuur voor bepaalde types van onderzoek bij te sturen (Ayris et al., 2018; Nosek et al., 2012, 2015).

1. Ondersteun replicatieonderzoek

Voor deze aanbeveling is het standpunt van de Koninklijke Nederlandse Akademie van Wetenschappen (2018) richtinggevend. Zij breekt een lans voor meer en systematisch replicatieonderzoek en formuleert haar aanbevelingen in de lijn van eerdere initiatieven door de Nederlandse Organisatie voor Wetenschappelijk Onderzoek en de Deutsche Forschungsgemeinschaft om te voorzien in aanzienlijke programmafinanciering voor replicatiestudies (Koninklijke Nederlandse Akademie van Wetenschappen, 2018, p. 66).

Replicatieonderzoek is belangrijk om repliceerbaarheidsproblemen op het spoor te komen en de vinger aan de pols te houden. Voor men theorieën en verklaringen ontwikkelt of aan toepassingen gaat denken, moet er namelijk eerst voldoende vertrouwen in de repliceerbaarheid van het betrokken empirische fenomeen worden opgebouwd (Zwaan et al., 2018). Reproduceerbaarheid is daarbij een noodzakelijke, maar geen voldoende voorwaarde. Het blijft ook aangewezen dat onafhankelijke onderzoekers met een nieuwe dataverzameling het empirische fenomeen kunnen herhalen (zie paragraaf 2.2.2). Repliceerbaarheid kan als de eerste externe kwaliteitscontrole voor onderzoek in een welbepaalde wetenschappelijke discipline worden beschouwd.

Replicatiestudies liggen voor de hand in methodologieonderwijs en masterproeven, maar ze moeten ook een prominentere plaats krijgen in het doorsneeonderzoek in de empirische menswetenschappen. Nu ligt de nadruk voor subsidieerbaarheid en publiceerbaarheid soms te veel op 'innovatie' en 'originaliteit', ten koste van 'methodologische disciplinering', 'controle' en 'consolidering' (De Boeck & Jeon, 2018; Koninklijke Nederlandse Akademie van Wetenschappen, 2018; Nosek et al., 2012, 2015).

Twee afsluitende opmerkingen:

(1) Replicatiestudies en uitgebreide replicatieprogramma's moeten uiteraard aan dezelfde methodologische standaarden voldoen als de oorspronkelijke studies. Bryan, Yeager en O'Brien (2019) toonden recent bijvoorbeeld aan dat een groot aantal *replicator degrees of freedom* voor een niet-replicatie kunnen zorgen. LeBel, Vanpaemel, Cheung en Campbell (2019) geven in dit verband interessante richtlijnen om replicatiestudies op een meer systematische manier te evalueren.

(2) In bepaalde onderzoeksdomeinen moeten we rekening houden met kleine en variabele effecten door de complexiteit en contextafhankelijkheid van de onderzochte empirische fenomenen (De Boeck & Jeon, 2018). In deze domeinen kunnen we geen of slechts een geringe repliceerbaarheid verwachten en kan het onderzoek beter worden gericht op het in kaart brengen van de complexiteit en contextafhankelijkheid dan op het zoeken naar algemene of conditionele

wetmatigheden. Bovendien is een geringe repliceerbaarheid op zichzelf een belangrijke bevinding. Wetenschappelijk onderzoek is namelijk ook functioneel als het aanzet tot bescheidenheid en als het overtrokken conclusies bij wetenschappers, journalisten en beleidsmakers tempert (Hughes, 2018; Koninklijke Nederlandse Akademie van Wetenschappen, 2018; Sumner et al., 2016).

2. Ondersteun open wetenschappelijk onderzoek

Voor deze aanbeveling is het recente advies van de League of European Research Universities (Ayris et al., 2018) richtinggevend. Dat advies bevat een grondig overzicht van de stappen die we kunnen zetten om naar meer openheid in het wetenschappelijk onderzoek te evolueren. Het advies wordt vergezeld van een handige checklist om tot concrete implementatie over te gaan.

Openheid in het wetenschappelijk onderzoek is belangrijk om replicatiestudies mogelijk te maken, om reproduceerbaarheid te onderzoeken en om QRPs te kunnen detecteren of te vermijden. Openheid is bovendien de beste remedie tegen repliceerbaarheidsproblemen die gekoppeld worden aan een vertekende en selectieve rapportering (zie paragrafen 3.3.2 en 3.3.3). Vele onderzoekers dragen openheid als een algemene wetenschappelijke waarde hoog in het vaandel en aanvaarden alleen overwegingen van vertrouwelijkheid en privacy om openheid over de data in te perken (De Boeck & Jeon, 2018; Merton, 1973; Nosek & Bar-Anan, 2012; Nosek et al., 2015).

Openheid is toepasbaar op alle fasen en aspecten van de empirische onderzoekscyclus (LeBel et al., 2018; Munafò et al., 2017; Van der Zee & Reich, 2018). Het is zelfs moeilijk om openheid voor een bepaald aspect te bepleiten en door te voeren (bv. open-access-publicaties) zonder de andere aspecten mee in beschouwing te nemen (All European Academies, 2018a 2018b). Een ideale non-profitmanier om een onderzoeksproject van A tot Z volgens de principes van open wetenschappelijk onderzoek te plannen en te beheren, biedt bijvoorbeeld het Open Science Framework.[11]

We geven hieronder kort nog enkele overwegingen om openheid in de verschillende fasen van de empirische onderzoekscyclus gestalte te geven: de ontologische en epistemologische positie, theorieën, hypothesen, predicties en geplande analyses, data, materialen en instrumenten, dataverwerking en/of statistische analyse, rapportering, *peerreview* en de verspreiding/beschikbaarheid van publicaties.

(1) Openheid over de ontologische en epistemologische positie van de onderzoekers wordt zelden, als een afzonderlijk aspect van openheid, met repliceerbaarheid in verband gebracht. Dit heeft waarschijnlijk te maken met het feit dat 'repli-

[11] https://osf.io/

ceerbaarheid' voornamelijk door onderzoekers die in een kritisch-realistische en natuurwetenschappelijke traditie werken als relevant wordt beschouwd. In sommige disciplines waar meer kwalitatieve en interpretatieve methoden gangbaar zijn, wordt echter wel verwacht dat men hierover open communiceert (Aguinis & Solarino, 2019; Morawski, 2019; Savin-Baden & Major, 2013).

(2) Openheid over theorieën, hypothesen, predicties en geplande analyses kan gebeuren in een preregistratie. Dat is een protocol waarin de onderzoeksplanning, uitvoering en analyse op voorhand worden vastgelegd, dat publiek toegankelijk is en dat van een gecertificeerde tijdstempel voorzien is.

Een bijzondere vorm van preregistratie is een *registered report*. Dit is een format dat meer en meer door wetenschappelijke tijdschriften wordt aangeboden en waarbij een manuscript op basis van een preregistratie en een begeleidende tekst wordt geëvalueerd, voorafgaand aan de dataverzameling. Als de evaluatie positief is, wordt het manuscript voorlopig aanvaard, onafhankelijk van de richting die de resultaten zullen uitgaan. Het gaat over een 'voorlopige' aanvaarding, omdat er achteraf ook nog een *review*ronde is voor het volledige manuscript (Chambers, 2019).

Eerste analyses wijzen uit dat er met *registered reports* minder statistische significante bevindingen worden gerapporteerd en dat de effecten kleiner zijn (Allen & Mehler, 2019; Warren, 2018; Wiseman, Watt, & Kornbrot, 2019). Dit kan positief worden geïnterpreteerd als het effect van *registered reports* op het vermijden van QRPs en het verminderde risico op vals positieven. We moeten afwachten wat verder onderzoek over hun effect zal uitwijzen, want de eerste effectanalyses zijn louter observationeel en correlationeel. Het zou kunnen dat onderzoekers het format van de *registered reports* gebruiken voor de studies waarbij zij zelf aan de afloop twijfelen (Allen & Mehler, 2019; Warren, 2018).

Voor de auteur die het format gebruikt, is er in elk geval het voordeel dat er bescherming is tegen CARKing (*Critiquing After the Results are Known*) door *reviewers* of *editors* (Nosek & Lakens, 2014). Omdat de *reviewers* en de *editors* in de eerste fase blind zijn gemaakt voor de uiteindelijke resultaten wordt vermeden dat er achteraf oneigenlijke methodekritiek of vitterij komt als de resultaten in hun ogen in een of ander opzicht tegenvallen.

(3) Openheid over data, materialen en instrumenten heeft van alle openheids-aspecten waarschijnlijk het grootste draagvlak. In dit Standpunt kunnen we dan ook zonder veel kanttekeningen het streven naar FAIR data onderschrijven (Wilkinson et al., 2016). FAIR data zijn data die aan de criteria *Findability*, *Accessibility*, *Interoperability* en *Reusability* voldoen (Wilkinson et al., 2016). Een open-databeleid is nodig om reproduceerbaarheid te onderzoeken, als noodzakelijke voorwaarde voor repliceerbaarheid. Daarbij is het belangrijk dat

er niet alleen getallen en symbolen beschikbaar zijn, maar dat de data ook goed gedocumenteerd zijn. Een open-databeleid houdt in dat er een waterdicht data-managementplan wordt opgemaakt en dat metadata ter beschikking worden gesteld (National Academies of Sciences, Engineering, and Medicine, 2019). Beleidsmakers zouden dit niet als een administratief onderdeel van het hele proces moeten zien – een rubriek méér in de projectaanvraag – maar ze moeten wel de nodige middelen en infrastructuur ter beschikking stellen om onderzoekers op te leiden en te ondersteunen bij dit onderdeel van het onderzoeksproces.

(4) Openheid over de dataverwerking en/of statistische analyse lijkt vanzelfsprekend, maar is dat niet. In veel onderzoek in de empirische menswetenschappen is de afstand tussen de verzamelde data en de gerapporteerde resultaten namelijk relatief groot vanwege verschillende filters, vormen van voorverwerking en complexe statistische analyses die op de data worden uitgevoerd. De beschikbaarheid van de programmacode en syntax is aan te bevelen boven menugestuurde analyses waarvan geen spoor wordt bijgehouden. Merk wel op dat reproduceerbaarheid van de dataverwerking en/of statistische analyse soms onherroepelijk wordt gecompromitteerd door compatibiliteitsproblemen tussen soorten hardware, *operating systems*, versies (National Academies of Sciences, Engineering, and Medicine, 2019)…

(5) Transparantie in de rapportering kan als een synoniem voor algemene openheid worden gebruikt en kan naar de verschillende aspecten van de empirische onderzoekscyclus verwijzen. Met 'transparantie in de rapportering' in enge zin gaat het over de eindrapportering in de vorm van een wetenschappelijk artikel, een boek(hoofdstuk) of een bijdrage in *proceedings*. De aanbeveling van transparantie in de rapportering is het meest rechtstreekse antwoord op de rapporteringsproblemen die in 3.3.2 en 3.3.3 aan de orde werden gesteld.

In het bijzonder voor tijdschriftredacties, financiers en wetenschappelijke verenigingen werden de TOP-richtlijnen opgesteld (*Transparency and Openness Promotion*; Nosek et al., 2015). Ze behelzen acht standaarden, elk met drie niveaus van gestrengheid. Redacties en bestuursraden kunnen kiezen welke standaarden ze wensen te implementeren en met welk niveau ze dat wensen te doen. Deze keuzemogelijkheid – achtmaal drie – biedt flexibiliteit om de TOP-richtlijnen te onderschrijven, rekening houdend met verschillen tussen disciplines, en zorgt tegelijk voor een gemeenschappelijk referentiekader (Nosek et al., 2015).

Het plaatsgebrek of de limiet in het aantal woorden van een klassiek wetenschappelijk artikel botste vroeger wel eens met de verwachte transparantie in de rapportering of werd soms ingeroepen als excuus voor het ontbreken van details in de beschrijving van materialen, procedures, methoden en technieken. Door de mogelijkheid die bijna elk tijdschrift nu aanbiedt om digitale bijlagen toe te voegen of digitale koppelingen naar eigen *repositories* te maken, is deze hindernis grotendeels uit de wereld geholpen (Imhoff, Smith, & van Zomeren, 2018).

Ten slotte willen we de aandacht vestigen op twee elementen van de rapporteringstransparantie die wel eens worden veronachtzaamd: de *Constraints on Generality* (COG) en de *Conflicts of Interest* (COI). COG verwijst naar de inferentiële beperking die inherent is aan onderzoek met menselijke deelnemers (Simons, Shoda, & Lindsay, 2017). Deze deelnemers zijn uit een welbepaalde populatie geselecteerd en op basis van het betreffende onderzoek zijn dan ook alleen uitspraken over die populatie te verantwoorden. Simons et al. (2017) stellen daarom voor om in de discussiesectie van al het empirische onderzoek dat gebaseerd is op menselijke deelnemers een *COG-statement* op te nemen dat de doelpopulatie voor de gerapporteerde resultaten aangeeft en verantwoordt. Door meer informatie te geven over de karakteristieken van de deelnemers aan het onderzoek wordt het opzetten van een replicatiestudie eenvoudiger. Door de doelpopulatie expliciet te maken en de conclusies van het onderzoek te beperken tot die doelpopulatie, wordt er bovendien ook een grotere bescheidenheid over de draagwijdte van het onderzoek ingebouwd (Simons et al., 2017).

De *COI-rapportering* is in de meeste tijdschriften standaard, maar wordt niet altijd nauwkeurig opgevolgd, zeker niet als het gaat over niet-financiële COI (bv. naambekendheid of reputatie door koppeling aan een bepaalde theorie of interventie die al dan niet door de empirische studie ondersteund wordt) of persoonlijke COI (bv. professionele of privérelaties met werknemers van organisaties of bedrijven die voor- of nadeel kunnen ondervinden bij de resultaten van de studie) (Lieb, Osten-Sacken, Stoffers-Winterling, Reiss, & Barth, 2016; Romero, 2018; Soares et al., 2019). Deze onderrapportering is betreurenswaardig omdat er naast de ethische dimensie bij COI ook een koppeling aan repliceerbaarheid is. Er is namelijk een verband tussen COI en de opwaartse vertekening van de effecten (Lieb et al., 2016; Soares et al., 2019). Voor Ioannidis (2005b) was COI zelfs een van de belangrijke motieven in zijn *Why Most Published Research Results Are False* (zie 3.1). Een eerlijke en open rapportering over financiële, niet-financiële en persoonlijke COI blijft dan ook aangewezen. Hoever men moet gaan in de rapportering van niet-financiële en persoonlijke COI verschilt van onderzoeksdomein tot onderzoeksdomein en vergt van de onderzoeker en de *editor* een persoonlijke afweging over de relevantie (Bero & Grundy, 2018; Ioannidis & Trepanowski, 2018).

(6) Over de openheid ten aanzien van *peerreview* en de *review*procedure is er minder eensgezindheid. Over de *open peerreview* wordt in de uitgeverswereld al langer gediscussieerd, onafhankelijk van de repliceerbaarheidsproblemen, omdat het gaat over een centrale schakel in de kwaliteitsbewaking van wetenschappelijke publicaties (DeCoursey, 1999, 2006). In het kader van repliceerbaarheidsproblemen zou *open peerreview* een belangrijke rol kunnen spelen omdat de huidige manier van werken (*closed review*, met *single* of *double blinding*) voor een *publicatiebias* en andere vertekeningen kan zorgen (Lee, Sugimoto, Zhang, & Cronin, 2013, en zie ook 3.1 en 3.3). In de TOP-richtlijnen van Nosek et al. (2015) komt het

echter niet voor. De laatste jaren wint *open peerreview* evenwel aan aanhang en wordt het beschouwd als een integraal deel van open wetenschappelijk onderzoek (Schmidt, Ross-Hellauer, Van Edig, & Moylan, 2018). Initiatieven zoals *Publons*[12] (sinds 2017 een deel van *Clarivate Analytics*) en de *open peerreview*procedure bij toonaangevende tijdschriften, zoals de *Proceedings of the Royal Society*, *Royal Society Open Science*, *BMJ*, *PeerJ*, *F1000Research* en *PLOS*, zijn in ieder geval een stimulans. Andere tijdschriften zullen volgen, zeker als het optioneel wordt gehouden.

(7) *Open access* staat hoog op de agenda van vele Europese universiteiten en het Europese onderzoeksbeleid (bv. met Plan S).[13] Openheid in de verspreiding en beschikbaarheid van publicaties is niet alleen belangrijk vanuit sociaal en democratisch oogpunt, zo ontsluit men bovendien de informatie die soms noodzakelijk is om replicatiestudies uit te kunnen voeren. Geheime documenten en hoge betaalmuren zijn de natuurlijke vijanden van open wetenschappelijk onderzoek, maar de concrete implementatie van algemene *open access* (of gedeeltelijke *open access*, zoals met Plan S voor de Europese onderzoeksprojecten tegen 2021) kan alleen maar lukken als tegelijk de evaluatiecriteria van het onderzoek en de onderzoekers worden hervormd en als er rekening wordt gehouden met de juridische, financiële en economische implicaties (All European Academies, 2018a 2018b).[14]

3. Ondersteun zorgvuldig wetenschappelijk onderzoek

Voor deze aanbeveling zijn de eindrapporten van The Academy of Medical Sciences in het Verenigd Koninkrijk (2015) en de National Academies of Sciences, Engineering, and Medicine in de Verenigde Staten (2019) richtinggevend. Zorgvuldig wetenschappelijk onderzoek richt zich rechtstreeks op de betrouwbaarheid van proefopzet, dataverzameling en data-analyse, die nodig zijn om onderzoek repliceerbaar te maken (zie 3.3.1). Bijzondere aandacht is nodig voor het verhogen van het onderscheidingsvermogen (Bishop, 2019; Button et al., 2013; Ioannidis, 2005b; Open Science Collaboration, 2015) en het vermijden van QRPs (John et al., 2012; Shrout & Rodgers, 2018; Nelson et al., 2018; Wicherts et al., 2016).

Het universitaire, nationale en internationale onderzoeksbeleid kan hier een belangrijke rol spelen door te vermijden dat men een klein aantal numerieke parameters gebruikt om onderzoekers aan te nemen, te beoordelen en te bevorderen, of

---

12 https://publons.com/about/home/

13 https://sparcopen.org/news/2018/coalition-european-funders-announces-plan-s/

14 Zie ook de visie van de Global Young Academy: Plan S is 'an invitation to contribute to shaping the research ecosystem and its impact on society as whole. At the same time, given the large room for possible interpretation and implementation, there is much concern that Plan S may not lead to the positive changes that we, as young scholars, think are necessary.' https://globalyoungacademy.net/ya-plan-s-statement/

om projectaanvragen te honoreren (Smaldino & McElreath, 2016). Op die manier kan de 'natuurlijke selectie' – bedoeld wordt: de onbedoelde selectieve vermenigvuldiging door de onderzoekscultuur – van kwalijke, maar volgens de geldende metrieken 'succesvolle' onderzoekspraktijken worden tegengegaan.

De ondersteuning van zorgvuldig wetenschappelijk onderzoek betekent ook dat men plaats vrijhoudt voor traag, degelijk en consoliderend onderzoek (Frith, 2020; Slow Science Academy, 2010), en voor grootschalig longitudinaal en collaboratief onderzoek (Nuijten, 2019; Uhlmann et al., 2019), naast snel en beleidsgericht onderzoek dat mikt op technologische innovatie.[15]

4. Ondersteun methodologieonderwijs

De drie aanbevelingen willen we flankeren met twee suggesties. De eerste suggestie staat zij aan zij met de derde aanbeveling: de ondersteuning van het methodologieonderwijs. Het gaat dan in de eerste plaats over het methodologieonderwijs aan de hogescholen en de universiteiten, maar ook al vroeger in de schoolloopbaan kunnen eerste stappen worden gezet.

Het is volgens ons belangrijk dat disciplinaire opleidingen zich niet alleen op theorieën, inhouden en praktische toepassingen richten, maar ook blijvend aandacht hebben voor de onderzoeksmethoden die werden en worden gebruikt om tot die theorieën, inhouden en praktische toepassingen te komen. Het methodologieonderwijs in de brede betekenis (dus ook de doctoraatsopleiding, onderzoeksstage en permanente vorming) is het eerste doorgeefluik van *responsible research practices*. Methodologieonderwijs met aandacht voor replicatieonderzoek, open wetenschap en zorgvuldigheid (zie aanbevelingen 1, 2 en 3) slaat drie vliegen in één klap (zie ook Ayris et al., 2018; Button, Chambers, Lawrence, & Munafò, 2019; Chopik, Bremner, Defever, & Keller, 2018; Frank & Saxe, 2012).

5. Ondersteun metaonderzoek.

De tweede flankerende suggestie betreft de ondersteuning van onderzoek *over onderzoek*. Dit zogenaamde metaonderzoek is onmiddellijk relevant voor de drie aanbevelingen: het is nodig om replicatiestudies te evalueren, om de impact van meer openheid in wetenschappelijk onderzoek te bestuderen en om te bepalen wat 'zorgvuldigheid' precies inhoudt.

---

[15] De nood aan een evenwichtige spreiding van onderzoeksmiddelen tussen enerzijds agendagedreven en onderzoeker-gedreven onderzoek en anderzijds fundamenteel en toegepast onderzoek werd ook bepleit in het Standpunt *Onderzoeker-gedreven wetenschap* https://www.kvab.be/nl/standpunten/onderzoeker-gedreven-wetenschap.

Het eindrapport van de National Academies of Sciences, Engineering, and Medicine in de Verenigde Staten (2019) verwacht veel heil van statistische meta-analyses, maar wij willen de ondersteuning van metaonderzoek breder opvatten. De meerderheid van de problemen die met repliceerbaarheid geassocieerd worden, is van niet-statistische aard of is niet met de beschikbaarheid van andere of complexere statistische technieken en modellen op te lossen (bv. QRPs, HARKen en *publicatiebias*). De ondersteuning van kritische, kwalitatieve en systematische *reviews*, eventueel aangevuld met meta-analyses, blijft dus wenselijk. Een samenwerking van inhoudelijke onderzoekers, methodologen en statistici kan daarbij vermijden dat de laatste twee uitsluitend als waakhonden en buitenwippers dienstdoen.

Ten slotte maakt dit Standpunt over repliceerbaarheid in de empirische menswetenschappen ook duidelijk dat het uitvoeren van wetenschappelijk onderzoek op zichzelf een vorm van menselijk gedrag in een welbepaalde maatschappelijke en economische context is (Hantula, 2019; Norris & O'Connor, 2019). Bovendien volstaat de empirische methode niet om zichzelf te onderzoeken, criteria en doelen te formuleren en zichzelf eventueel bij te sturen (Morawski, 2019). 'Metaonderzoek' wordt daarom het best breed en interdisciplinair gedefinieerd, waarbij zowel de psychologische als de sociale, politieke, economische, juridische, filosofische en ethische dimensies van wetenschappelijk onderzoek worden betrokken (Gholson, Shadish, Neimeyer, & Houts, 1989; Ioannidis, 2018). Metaonderzoek wordt zo de plaats bij uitstek waar de menswetenschappelijke disciplines elkaar ontmoeten.

Conclusie

De repliceerbaarheidsstorm die in het begin van de 21ste eeuw opstak en momenteel door de empirische menswetenschappen raast, is nog lang niet gaan liggen. Met dit Standpunt hebben we geprobeerd te beargumenteren dat we er een dam tegen kunnen opwerpen en zo de storm bedwingen: door ondersteuning van replicatiestudies, open wetenschap en zorgvuldig wetenschappelijk onderzoek, geflankeerd door gedegen methodologieonderwijs en metaonderzoek. Als we als wetenschappers maatschappelijk respect willen blijven verdienen en als we wensen dat een onderzoeksgebaseerde aanbeveling niet als een mening naast vele andere meningen kan worden weggezet, dan is het hoog tijd om een nieuwe wind te doen waaien.

Referenties

Aczel, B., Szaszi, B., Sarafoglou, A., Kekecs, Z., Kucharský, Š., Benjamin, D., ... Wagenmakers, E.-J. (2019). A consensus-based transparency checklist. *Nature Human Behaviour.* doi:10.1038/s41562-019-0772-6

Agnoli, F., Wicherts, J. M., Veldkamp, C. L. S., Albiero, P., & Cubelli, R. (2017). Questionable research practices among Italian research psychologists. *PLoS ONE, 12*(3), e0172792. doi:10.1371/journal.pone.0172792

Aguinis, H., & Solarino, A. M. (2019). Transparency and replicability in qualitative research: The case of interviews with elite informants. *Strategic Management Journal, 40,* 1291-1315. doi:10.1002/smj.3015

All European Academies (2018a). *ALLEA response to Plan S.* Online https://allea.org/portfolio-item/allea-response-to-plan-s/

All European Academies (2018b). *Ethical aspects of open access: A windy road. Workshop report.* Online https://allea.org/ethical-aspects-of-open-access-a-windy-road/

All European Academies (2018c). *Loss of trust? Loss of trustworthiness? Truth and expertise today. ALLEA Discussion Paper 1.* Online https://allea.org/portfolio-item/loss-of-trust-loss-of-trustworthiness-truth-and-expertise-today/

All European Academies (2019a). *Trust within science: Dynamics and norms of knowledge production. ALLEA Discussion Paper 2.* Online https://allea.org/portfolio-item/trust-within-science-dynamics-and-norms-of-knowledge-production/

All European Academies (2019b). *Trust in science and changing landscapes of communication. ALLEA Discussion Paper 3.* Online https://allea.org/portfolio-item/trust-in-science-and-changing-landscapes-of-communication/

Altman, D. G., Moher, D., & Schulz, K. F. (2017). Harms of outcome switching in reports of randomised trials: CONSORT perspective. *BMJ, 356,* j396. doi:10.1136/bmj.j396

Armitage, P., McPherson, C. K., & Rowe, B. C. (1969). Repeated significance tests on accumulating data. *Journal of the Royal Statistical Society, 132,* 235-244. doi:10.2307/2343787

Artino, A. R., Jr., Driessen, E. W., & Maggio, L. A. (2019). Ethical shades of gray: International frequency of scientific misconduct and questionable research practices in health professions education. *Academic Medicine, 94,* 76-84. doi:10.1097/ACM.0000000000002412

Ayris, P., López de San Román, A., Maes, K., & Labastida, I. (2018). *Open science and its role in universities: A roadmap for cultural change.* LERU Advice paper 24. Leuven, BE: League of European Research Universities.

Baker, M. (2016). 1,500 scientists lift the lid on reproducibility. *Nature, 533*(7604), 452-454. doi:10.1038/533452a

Bakker, M., & Wicherts, J. M. (2014a). Outlier removal and the relation with reporting errors and quality of psychological research. *PLoS ONE, 9*(7), e103360. doi:10.1371/journal.pone.0103360

Bakker, M., & Wicherts, J. M. (2014b). Outlier removal, sum scores, and the inflation of the Type I error rate in independent samples *t* tests: The power of alternatives and recommendations. *Psychological Methods, 19,* 409-427. doi: 10.1037/met0000014

Banks, G. C., Rogelberg, S. G., Woznyj, H. M., Landis, R. S., & Rupp, D. E. (2016). Editorial: Evidence on questionable research practices: The good, the bad, and the ugly. *Journal of Business and Psychology, 31*(3), 323-338. https://doi.org/10.1007/s10869-016-9456-7

Barnett, V., & Lewis, T. (1994). *Outliers in statistical data* (3rd ed.). New York, NY: Wiley.

Barsalou, L. W. (2016). Situated conceptualization offers a theoretical account of social priming. *Current Opinion in Psychology 12,* 6-11. doi:10.1016/j.copsyc.2016.04.009

Belia, S., Fidler, F., Williams, J., & Cumming, G. (2005). Researchers misunderstand confidence intervals and standard error bars. *Psychological Methods, 10,* 389-396. doi:10.1037/1082-989X.10.4.389

Bem, D. J. (2004). Writing the empirical journal article. In J. M. Darley, M. P. Zanna, & H. L. Roediger III (Eds.), *The compleat academic: A practical guide for the beginning social scientist* (2nd ed., pp. 185-219). Washington, DC: American Psychological Association.

Bem, D. J. (2011). Feeling the future: experimental evidence for anomalous retroactive influences on cognition and affect. *Journal of Personality and Social Psychology, 100,* 407-425. doi:10.1037/a0021524

Bem, D., Tressoldi, P., Rabeyron, T., & Duggan, M. (2015/2016). Feeling the future: A meta-analysis of 90 experiments on the anomalous anticipation of random future events. *F1000Research, 4,* 1188. Version 1, 2015 Oct 30, and Version 2, 2016 Jan 29 doi:10.12688/f1000research.7177.2

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society, Series B, 57,* 289-300. doi:10.1111/j.2517-6161.1995.tb02031.x

Bennett, C. M., Baird, A. A., Miller, M. B., & Wolford, G. L. (2009). *Neural correlates of interspecies perspective taking in the post-mortem Atlantic Salmon: An argument for multiple comparisons correction.* Poster presented at the 15th Annual Meeting of the Organization for Human Brain Mapping. San Francisco, CA. Poster available at http://users.stat.umn.edu/~corbett/classes/5303/Bennett-Salmon-2009.pdf

Bennett, C. M., Miller, M. B., & Wolford, G. L. (2009). The principled control of false positives in neuroimaging. *Social Cognitive and Affective Neuroscience, 4,* 417-422. doi:10.1093/scan/nsp053.

Bergmann, C., Tsuji, S., Piccinini, P. E., Lewis, M. L., Braginsky, M., Frank, M. C., & Cristia, A. (2018). Promoting replicability in developmental research through meta-analyses: Insights from language acquisition research. *Child Development, 89,* 1996-2009. doi:10.1111/cdev.13079

Bero, L., & Grundy, Q. (2018). Conflicts of interest in nutrition research. *Journal of the American Medical Association, 320,* 93-94. doi:10.1001/jama.2018.5662

Biswal, B. B., et al. (2010). Toward discovery science of human brain function. *Proceedings of the National Academy of Sciences of the United States of America, 107*(10), 4734-4739. doi:10.1073/pnas.0911855107

Billiet, J., Opgenhaffen, M., Pattyn, B., & Van Aelst, P. (2018). *De strijd om de waarheid. Over nepnieuws en desinformatie in de digitale mediawereld: KVAB Standpunt 62.* Brussel: KVAB.

Bland, J. M., & Altman, D. G. (1996a). The use of transformation when comparing two means. *BMJ, 312,* 1153.

Bland, J. M., & Altman, D. G. (1996b). Transforming data. *BMJ, 312,* 770.

Bollen, K., Cacioppo, J. T., Kaplan, R. M., Knosnick, J. A., & Olds, J. L. (2015). *Social, behavioral, and economic sciences perspectives on robust and reliable science.* Report of the Subcommittee on Replicability in Science Advisory Committee to the National Science Foundation Directorate for Social, Behavioral, and Economic Sciences. https://www.nsf.gov/sbe/AC_Materials/SBE_Robust_and_Reliable_Research_Report.pdf

Bollen, K. A., & Jackman, R. W. (1990). Regression diagnostics: An expository treatment of outliers and influential cases. In J. Fox & J. S. Long (Eds.), *Modern methods of data analysis* (pp. 257-291). Newbury Park, CA: Sage.

Borenstein, M., Hedges, L. V., Higgins, J. P. T., & Rothstein, H. R. (2009). *Introduction to meta-analysis.* Chichester, UK: Wiley.

Borghi, A. M., & Fini, C. (2019). Theories and explanations in psychology. *Frontiers in Psychology, 10,* 1-3. doi:10.3389/fpsyg.2019.00958

Bösch, H., Steinkamp, F., & Boller E. (2006). Examining psychokinesis: The interaction of human intention with random number generators: A meta-analysis. *Psychological Bulletin, 132,* 497-523. doi:10.1037/0033-2909.132.4.497

Bower, G. H., & Mayer, J. D. (1985). Failure to replicate mood-dependent retrieval. *Bulletin of the Psychonomic Society, 23,* 39-42. doi:10.3758/BF03329773

Box, G. E. P. (1953). Non-normality and tests on variances. *Biometrika, 40,* 318-335.

Box, G. E. P., Leonard, T., & Wu, C.-F. (Eds.) (1983). *Scientific inference, data analysis, and robustness.* New York, NY: Academic Press.

Brainard, J. (2018). Rethinking retractions. *Science, 362*(6413), 390-393. doi:10.1126/science.362.6413.390

Brito, R., & Rodríguez-Navarro, A. (2019). Evaluating research and researchers by the journal impact factor: Is it better than coin flipping? *Journal of Informetrics, 13,* 314-324. doi:10.1016/j.joi.2019.01.009

Bronstein, R. F. (1990). Publication politics, experimenter bias and the replication process in social science research. *Journal of Social Behavior and Personality, 5,* 71-81.

Bryan, C. J., Yeager, D. S., & O'Brien, J. M. (2019). Replicator degrees of freedom allow publication of misleading failures to replicate. *Proceedings of the National Academy of Sciences, 116,* 25535-25545. doi:10.1073/pnas.1910951116

Bucci, E. M. (2019). On zombie papers. *Cell, Death & Disease, 10,* 189. doi:10.1038/s41419-019-1450-3

Button, K. S., Chambers, C. D., Lawrence, N., & Munafò, M. R. (2019). Grassroots training for reproducible science: A consortium-based approach to the empirical dissertation. *Psychology Learning & Teaching*, 147572571985765. doi:10.1177/1475725719857659

Button, K. S., Ioannidis, J. P. A., Mokrysz, C., Nosek, B. A., Flint, J., Robinson, E. S. J., & Munafò, M. R. (2013). Power failure: Why small sample size undermines the reliability of neuroscience. *Nature Reviews Neuroscience, 14,* 365-376. doi:10.1038/nrn3475

Camerer, C. F., Dreber, A., Forsell, E., Ho, T.-H., Huber, J., Johannesson, M., …, Wu, H. (2016). Evaluating replicability of laboratory experiments in economics. *Science, 351*(6280), 1433-1436. doi:10.1126/science.aaf0918

Camerer, C.F., et al. (2018). Evaluating the replicability of social science experiments in Nature and Science between 2010 and 2015. *Nature Human Behavior, 2,* 637-644. doi:10.1038/s41562-018-0399-z

Caon, M. (2017). Gaming the impact factor: where who cites what, whom and when. *Australasian Physical & Engineering Sciences in Medicine, 40,* 273-276. doi:10.1007/s13246-017-0547-1

Carnap, R., (1967). *The logical structure of the world* [Der logische Aufbau der Welt]. Berkeley, CA: University of California Press. (Original work published 1928)

Carney, D. R., Cuddy, A. J. C., & Yap, A. J. (2010). Power posing: Brief nonverbal displays affect neuroendocrine levels and risk tolerance. *Psychological Science, 21,* 1363-1368. doi:10.1177/0956797610383437

Chambers, C. (2019). What's next for Registered Reports? *Nature, 573*(7773), 187-189. doi:10.1038/d41586-019-02674-6

Chan, A., Hróbjartsson, A., Haahr, M. T., Gøtzsche, P. C., & Altman, D. G. (2004). Empirical evidence for selective reporting of outcomes in randomized trials: Comparison of protocols to published articles. *Journal of the American Medical Association, 291,* 2457-2465. doi:10.1001/jama.291.20.2457

Chartier, C., Kline, M., McCarthy, R., Nuijten, M., Dunleavy, D. J., & Ledgerwood, A. (2018, November 30). The cooperative revolution is making psychological science better. *Association for Psychological Science Observer,* online https://www.psychologicalscience.org/observer/the-cooperative-revolution-is-making-psychological-science-better

Chhin, C. S., Taylor, K. A., & Wei, W. S. (2018). Supporting a culture of replication: An examination of education and special education research grants funded by the Institute of Education Sciences. *Educational Researcher, 47,* 594-605. doi:10.3102/0013189X18788047

Chopik, W. J., Bremner, R. H., Defever, A. M., & Keller, V. N. (2018). How (and whether) to teach undergraduates about the replication crisis in psychological science. *Teaching of Psychology, 45,* 158-163. doi:10.1177/0098628318762900

Christensen, G., & Miguel, E. (2018). Transparency, reproducibility, and the credibility of economics research. *Journal of Economic Literature, 56,* 920-980. doi:10.1257/jel.20171350

Claesen, A., Gomes, S. L. B. T., Tuerlinckx, F., & Vanpaemel, W. (2019, May 12). Preregistration: Comparing dream to reality. *PsyArXiv* https://doi.org/10.31234/osf.io/d8wex

Cohen, J. (1962). The statistical power of abnormal–social psychological research: A review. *Journal of Abnormal and Social Psychology, 65,* 145-153. doi:10.1037/h0045186

Cohen, J. (1969). *Statistical power analysis for the behavioral sciences.* San Diego, CA: Academic Press.

Collins, H. M. (1985). *Changing order: Replication and induction in scientific practice.* London, UK: Sage.

Cooper, H., DeNeve, K., & Charlton, K. (1997). Finding the missing science: The fate of studies submitted for review by a human subjects committee. *Psychological Methods, 2,* 447-452. doi:10.1037/1082-989X.2.4.447

Cooper, M. M. (2018). The replication crisis and chemistry education research. *Journal of Chemical Education, 95,* 1-2. doi:10.1021/acs.jchemed.7b00907

Coyne, J. C. (2016). Replication initiatives will not salvage the trustworthiness of psychology. *BMC Psychology, 4,* 28. doi:10.1186/s40359016-0134-3

Credé, M. (2019). A negative effect of a contractive pose is not evidence for the positive effect of an expansive pose: Comment on Cuddy, Schultz, and Fosse (2018). *Meta-Psychology, 3,* 1723. doi:10.15626/MP.2019.1723

Cronbach, L. J. (1982). *Designing evaluations of educational and social programs*. San Francisco, CA: Jossey-Bass.

Cuddy, A. (2015). *Presence: Bringing your boldest self to your biggest challenges.* Boston, MA: Little, Brown, & Company.

Cuddy, A. J. C., Schultz, S. J., & Fosse, N. E. (2018). P-curving a more comprehensive body of research on postural feedback reveals clear evidential value for power-posing effect: Reply to Simmons and Simonsohn (2017). *Psychological Science, 29,* 656-666. doi:10.1177/0956797617746749

Cumming, G., & Maillardet, R. (2006). Confidence intervals and replication: Where will the next mean fall? Psychological Methods, 11, 217-227. doi:10.1037/1082-989X.11.3.217

Darley, J. M., Zanna, M. P., & Roediger III, H. L. (Eds.) (2004). *The compleat academic: A practical guide for the beginning social scientist* (2nd ed.). Washington, DC: American Psychological Association.

De Boeck, P., & Jeon, M. (2018). Perceived crisis and reforms: Issues, explanations, and remedies. *Psychological Bulletin, 144,* 757-777. doi:10.1037/bul0000154

DeCoursey, T. (1999). Pros and cons of open peer review. *Nature Neuroscience, 2*(3), 197-198. doi:10.1038/6295

DeCoursey, T. (2006). Perspective: The pros and cons of open peer review. Should authors be told who their reviewers are? *Nature*, online https://www.nature.com/nature/peerreview/debate/nature04991.html doi:10.1038/nature04991

de Groot, A. D. (1961). *Methodologie: Grondslagen van onderzoek en denken in de gedragswetenschappen.* Den Haag, NL: Mouton.

Dickersin, K., Chan, S., Chalmers, T. C., Sacks, H. S., & Smith Jr., H. (1987). Publication bias and clinical trials. *Controlled Clinical Trials, 8,* 343-353. doi:10.1016/0197-2456(87)90155-3

Dickersin, K., Min, Y. I., & Meinert, C. L. (1992). Factors influencing publication of research results. Follow-up of applications submitted to two institutional review boards. *Journal of the American Medical Association, 267,* 374-378. doi:10.1001/jama.1992.03480030052036

Dominus, S. (2017, October 18). When the revolution came for Amy Cuddy. *The New York Times Magazine*, online https://www.nytimes.com/2017/10/18/magazine/when-the-revolution-came-for-amy-cuddy.html

Dumas-Mallet, E., Button, K. S., Boraud, T., Gonon, F., & Munafò, M. R. (2017). Low statistical power in biomedical science: a review of three human research domains. *Royal Society Open Science, 4*(2), 160254. doi: 10.1098/rsos.160254

Earp, B. D., & Trafimow, D. (2015). Replication, falsification, and the crisis of confidence in social psychology. *Frontiers in Psychology, 6,* 621. doi:10.3389/fpsyg.2015.00621

Ebersole, C. R., Atherton, O. E., Belanger, A. L., Skulborstad, H. M., Allen, J., Banks, J. B., … Nosek, B. A. (2016). Many Labs 3: Evaluating participant pool quality across the academic semester via replication. *Journal of Experimental Social Psychology, 67,* 68-82. 10.1016/j.jesp.2015.10.012

Ellemers, N. (2013). Connecting the dots: Mobilizing theory to reveal the big picture in social psychology (and why we should do this). *European Journal of Social Psychology, 43,* 1-8. doi:10.1002/ejsp.1932

Elmenreich, W., Moll, P., Theuermann, S., & Lux, M. (2018). Making computer science results reproducible: A case study using Gradle and Docker. *PeerJ Preprints 6*, e27082v1. doi:10.7287/peerj.preprints.27082v1

Elsesser, K. (2018, April 3). Power posing is back: Amy Cuddy successfully refutes criticism. *Forbes,* online https://www.forbes.com/sites/kimelsesser/2018/04/03/power-posing-is-back-amy-cuddy-successfully-refutes-criticism/#31688c4f3b8e

Engber, D. (2017). Daryl Bem proved ESP Is real: Which means science is broken. https://slate.com/health-and-science/2017/06/daryl-bem-proved-esp-is-real-showed-science-is-broken.html

Falk-Krzesinski, H. J., & Tobin, S. C. (2015). How do I review thee? Let me count the ways: A comparison of research grant proposal review criteria across US federal funding agencies. *Journal of Research Administration, 46,* 79-94.

Fang, F. C., & Casadevall, A. (2011). Retracted science and the retraction index. *Infection and Immunity, 79,* 3855-3859. doi: 10.1128/IAI.05661-11

Ferguson, C. J. (2015). "Everybody knows psychology is not a real science": Public perceptions of psychology and how we can improve our relationship with policymakers, the scientific community, and the general public. *American Psychologist, 70,* 527-542. doi: 10.1037/a0039405.

Ferguson, C. J., Brown, J. M., & Torres, A. V. (2018). Education or indoctrination? The accuracy of introductory psychology textbooks in covering controversial topics and urban legends about psychology. *Current Psychology, 37,* 574-582. doi:10.1007/s12144-016-9539-7

Feyerabend, P. (1975). *Against method: Outline of an anarchistic theory of knowledge.* London, UK: New Left Books.

Fisher, C. R. (2014). A pedagogic exploration of researcher degrees of freedom. *Spreadsheets in Education, 7*(1), Article 1. Available at http://epublications.bond.edu.au/ejsie/vol7/iss1/1

Fisher, R. A. (1935). *The design of experiments.* Oxford, UK: Oliver & Boyd.

Fiske, S. T. (2016, November). A call to change science's culture of shaming. *Association for Psychological Science Observer,* online https://www.psychologicalscience.org/observer/acall-to-change-sciences-culture-of-shaming

Flake, J. K., & Fried, E. I. (2019, January 17). *Measurement schmeasurement: Questionable measurement practices and how to avoid them.* PsyArXiv preprint https://doi.org/10.31234/osf.io/hs7wm

Francis, G. (2012). Publication bias and the failure of replication in experimental psychology. *Psychonomic Bulletin & Review, 19,* 975-991. doi:10.3758/s13423-012-0322-y

Francis, G. (2014). The frequency of excess success for articles in Psychological Science. *Psychonomic Bulletin & Review, 21,* 1180-1187. doi:10.3758/s13423-014-0601-x

Franco, A., Malhotra, N., & Simonovits, G. (2014). Publication bias in the social sciences: Unlocking the file drawer. *Science, 345*(6203), 1502-1505. doi:10.1126/science.1255484

Frank, M. C., & Saxe, R. (2012). Teaching Replication. *Perspectives on Psychological Science, 7,* 600-604. doi:10.1177/1745691612460686

French, C. (2012, March 15). Precognition studies and the curse of the failed replications. *The Guardian*, online https://www.theguardian.com/science/2012/mar/15/precognition-studies-curse-failed-replications

Frith, U. (2020). Fast lane to slow science. *Trends in Cognitive Sciences, 24,* 1-2. doi:10.1016/j.tics.2019.10.007

Funk, C., Gottfried, J., & Mitchell, A. (2017, September 20). Science news and information today. *Pew Research Center,* September 20. Online http://www.journalism.org/2017/09/20/science-news-and-information-today

Galak, J., LeBoeuf, R. A., Nelson, L. D., & Simmons, J. P. (2012). Correcting the past: Failures to replicate psi. *Journal of Personality and Social Psychology, 103,* 933-948. doi:10.1037/a0029709

Gartlehner, G., et al. (2016). Comparative benefits and harms of antidepressant, psychological, complementary, and exercise treatments for major depression: An evidence report for a clinical practice guideline from the American College of Physicians. *Annals of Internal Medicine, 164,* 331-341. doi:10.7326/M15-1813

Gholson, B., Shadish, W. R., Jr., Neimeyer, R. A., & Houts, A. C. (Eds.). (1989). *Psychology of science: Contributions to metascience.* Cambridge, UK: Cambridge University Press.

Gilbert, D. T., King, G., Pettigrew, S., & Wilson, T. D. (2016). Comment on "Estimating the reproducibility of psychological science." *Science, 351,* 1037–1038. http://dx.doi.org/10.1126/science.aad7243

Gillies, D. (1998). The Duhem thesis and the Quine thesis. In M. Curd & J. A. Cover (Ed.), *Philosophy of science: The central issues* (pp. 302–319). New York, NY: Norton.

Goldacre, B. (2016). Make journals report clinical trials properly. *Nature, 530*(7588), 7. doi:10.1038/530007a

Greenwald, A. G. (1975). Consequences of prejudice against the null hypothesis. *Psychological Bulletin, 82,* 1–20. doi:10.1037/h0076157

Grimes, D. R., Bauch, C. T., & Ioannidis, J. P. A. (2018). Modelling science trustworthiness under publish or perish pressure. *Royal Society Open Science, 5,* 171511. doi:10.1098/rsos.171511

Gunsalus, C. K., & Robinson, A. D. (2018). Nine pitfalls of research misconduct. *Nature, 557,* 297–299. doi:10.1038/d41586-018-05145-6

Hansen, L. P., & Sargent, T. J. (2008). *Robustness.* Princeton, NJ: Princeton University Press.

Hantula, D. A. (2019). Editorial: Replication and reliability in behavior science and behavior analysis: A call for a conversation. *Perspectives on Behavior Science, 42,* 1–11. doi:10.1007/s40614-019-00194-2

Harding, S. (1976). *Can theories be refuted? Essays on the Duhem-Quine thesis.* Dordrecht, NL: D. Reidel Publishing Company.

Harvey, C. R., Liu, Y., & Zhu, H. (2016). . . . and the cross-section of expected returns. *Review of Financial Studies, 29,* 5–68. doi:10.1093/rfs/hhv059

Head, M. L., Holman, L., Lanfear, R., Kahn, A. T., & Jennions, M. D. (2015). The extent and consequences of *p*-hacking in science. *PLoS Biology, 13*(3), e1002106. doi:10.1371/journal.pbio.1002106

Hedges, L. V., & Schauer, J. M. (2019a). Consistency of effects is important in replication: Rejoinder to Mathur and VanderWeele (2019). *Psychological Methods, 24,* 576–577. doi:10.1037/met0000237

Hedges, L. V., & Schauer, J. M. (2019b). More than one replication study is needed for unambiguous tests of replication. *Journal of Educational and Behavioral Statistics, 44,* 543–570. doi:10.3102/1076998619852953

Hedges, L. V., & Schauer, J. M. (2019c). Statistical analyses for studying replication: Meta-analytic perspectives. *Psychological Methods, 24,* 557–570. doi:10.1037/met0000189

Hengartner, M. P. (2018). Raising awareness for the replication crisis in clinical psychology by focusing on inconsistencies in psychotherapy research: How much can we rely on published findings from efficacy trials? *Frontiers in Psychology, 9,* 256. doi:10.3389/fpsyg.2018.00256

Hilgard, J., & Jamieson, K. H. (2017). Science as "broken" versus science as "self-correcting": How retractions and peer-review problems are exploited to attack science. In K. H. Jamieson, D. Kahan, & D. A. Scheufele (Eds.), *The Oxford handbook of the science of science communication* (pp. 85–92). New York, NY: Oxford University Press.

Hollenbeck, J. R., & Wright, P. M. (2017). Harking, sharking, and tharking: Making the case for post hoc analysis of scientific data. *Journal of Management, 43,* 5–18. doi:10.1177/0149206316679487

Hopf, H., Krief, A., Mehta, G., & Matlin, S. A. (2019). Fake science and the knowledge crisis: Ignorance can be fatal. *Royal Society Open Science, 6,* 190161. doi:10.1098/rsos.190161

Hoyningen-Huene, P. (1987). Context of discovery and context of justification. *Studies in History and Philosophy of Science Part A, 18,* 501–515. doi:10.1016/0039-3681(87)90005-7

Hughes, B. M. (2018). *Psychology in crisis.* London, UK: Palgrave.

Imhoff, R., Smith, J., & van Zomeren, M. (2018). Editorial: Opening up to openness. *European Journal of Social Psychology, 48,* 1–3. doi:10.1002/ejsp.2359

Ioannidis, J. P. (2005a). Contradicted and initially stronger effects in highly cited clinical research. *Journal of the American Medical Association, 294,* 218–228. doi:10.1001/jama.294.2.218

Ioannidis, J. P. (2005b). Why most published research findings are false. *PLoS Medicine, 2*(8), e124. doi:10.1371/journal.pmed.0020124

Ioannidis, J. P. A. (2012). Why science is not necessarily self-correcting. *Perspectives on Psychological Science, 7,* 645–654. doi:10.1177/1745691612464056

Ioannidis, J. P. A. (2018). Meta-research: Why research on research matters. *PLoS Biology, 16,* e2005468. doi: 10.1371/journal.pbio.2005468.

Ioannidis, J. P., Munafò, M. R., Fusar-Poli, P., Nosek, B. A., & David, S. P. (2014). Publication and other reporting biases in cognitive sciences: detection, prevalence, and prevention. *Trends in Cognitive Sciences, 18,* 235–241. doi:10.1016/j.tics.2014.02.010

Ioannidis, J. P. A., Stanley, T. D., & Doucouliagos, H. (2017). The power of bias in economics research. *The Economic Journal, 127,* F236–F265. doi:10.1111/ecoj.1246

Ioannidis, J. P. A., & Trepanowski, J. F. (2018). Conflict of interest in nutrition research: Reply. *Journal of the American Medical Association, 320,* 94–95. doi: 10.1001/jama.2018.5678

Ipsos (2019, August). *Global trust in professions: Who do global citizens trust?* https://www.ipsos.com/sites/default/files/ct/news/documents/2019-09/global-trust-in-professions-trust-worthiness-index-2019.pdf

Jamieson, K. H. (2018). Crisis or self-correction: Rethinking media narratives about the well-being of science. *Proceedings of the National Academy of Sciences of the United States of America, 115,* 2620–2627. doi:10.1073/pnas.1708276114

John, L., Loewenstein, G. F., & Prelec, D. (2012). Measuring the prevalence of questionable research practices with incentives for truth-telling. *Psychological Science, 23,* 524–532. doi:10.1177/0956797611430953

Jones, C. W., Keil, L. G., Holland, W. C., Caughey, M. C., & Platts-Mills, T. F. (2015). Comparison of registered and published outcomes in randomized controlled trials: A systematic review. *BMC Medicine, 13,* 282. doi: 10.1186/s12916-015-0520-3

Kerr, N. L. (1998). HARKing: Hypothesizing after the results are known. *Personality and Social Psychology Review, 2,* 196–217. doi:10.1207/s15327957pspr0203_4

Kiai, A. (2019). To protect credibility in science, banish "publish or perish". *Nature Human Behaviour, 3,* 1017–1018. doi:10.1038/s41562-019-0741-0

Klein, R. A., Ratliff, K. A., Vianello, M., Adams, R. B., Jr., Bahník, Š., Bernstein, M. J., ... Nosek, B. A. (2014). Investigating variation in replicability: A "Many Labs" replication project. *Social Psychology, 45,* 142–152. doi:10.1027/1864-9335/a000178

Klein, R. A., Vianello, M., Hasselman, F., Adams, B. G., Adams, R. B., Jr., Alper, S., … Nosek, B. A. (2018). Many Labs 2: Investigating variation in replicability across samples and settings. *Advances in Methods and Practices in Psychological Science, 1,* 443–490. doi:10.1177/2515245918810225

Klein, R. A., Cook, C. L., Ebersole, C. R., Vitiello, C. A., Nosek, B. A., Chartier, C. R., … Ratliff, K. A. (2019, December 11). *Many Labs 4: Failure to replicate mortality salience effect with and without original author involvement.* Preprint doi:10.31234/osf.io/vef2c

Klein, S. B. (2014). What can recent replication failures tell us about the theoretical commitments of psychology? *Theory & Psychology, 24,* 326–338. doi:10.1177/0959354314529616

Koninklijke Nederlandse Akademie van Wetenschappen (2018). *Replication studies: Improving reproducibility in the empirical sciences.* Amsterdam, NL: Koninklijke Nederlandse Akademie van Wetenschappen.

Krishna, A., & Peter, S. M. (2018). Questionable research practices in student final theses: Prevalence, attitudes, and the role of the supervisor's perceived attitudes. *PLoS ONE, 13*(8), e0203470. doi:10.1371/journal.pone.0203470

Kuhn, T. (1962). *The structure of scientific revolutions.* Chicago, IL: University of Chicago Press.

Kwa, C. (2018). *What is truth? A new philosophy of the sciences and the humanities.* Amsterdam, NL: Boom.

Lakatos, I., & Musgrave, A. (Eds.) (1970). *Criticism and the growth of knowledge.* Cambridge, UK: Cambridge University Press.

Lawn, J. (2011). Brian Wansink. *Food Management, 46*(7), 36-38.

LeBel, E. P., McCarthy, R. J., Earp, B. D., Elson, M., & Vanpaemel, W. (2018). A unified framework to quantify the credibility of scientific findings. *Advances in Methods and Practices in Psychological Science, 1,* 389-402. doi:10.1177/2515245918787489

LeBel, E. P., Vanpaemel, W., Cheung, I., & Campbell, L. (2019). A brief guide to evaluate replications. *Meta-psychology, 3,* 843. doi:10.15626/MP.2018.843

Lee, C. J., Sugimoto, C. R., Zhang, G., & Cronin, B. (2013). Bias in peer review. *Journal of the American Society for Information Science and Technology, 64,* 2-17. doi:10.1002/asi.22784

Lee, S. M. (2018, September 20). Cornell just found Brian Wansink guilty of scientific misconduct and he has resigned. *Buzzfeed News*, online https://web.archive.org/web/20180920200402/https://www.buzzfeednews.com/article/stephaniemlee/brian-wansink-retired-cornell

Leezenberg, M., & de Vries, G. (2017). *Wetenschapsfilosofie voor geestes-wetenschappen* (3de ed.). Amsterdam, NL: Amsterdam University Press.

Levelt, W. J., Drenth, P., & Noort, E. (2012). *Falende wetenschap: De frauduleuze onderzoekspraktijken van sociaal-psycholoog Diederik Stapel.* Tilburg University, Universiteit Amsterdam en Universiteit Groningen. Beschikbaar op https://onderwijsbrabant.nl/sites/default/files/eindrapport_stapel_nov_2012.pdf

Leydesdorff, L., Bornmann, L., Comins, J. A., & Milojevic, S. (2016). Citations: Indicators of quality? The impact fallacy. *Frontiers in Research Metrics and Analytics, 1,* 1. 10.3389/frma.2016.00001

Lieb, K., Osten-Sacken, J. V. D., Stoffers-Winterling, J., Reiss, N., & Barth, J. (2016). Conflicts of interest and spin in reviews of psychological therapies: A systematic review. *BMJ Open, 6,* e010606. doi: 10.1136/bmjopen-2015-010606

Lilienfeld, S. O. (2017). Psychology's replication crisis and the grant culture: Righting the ship. *Perspectives on Psychological Science, 12,* 660-664. doi:10.1177/1745691616687745

Lilienfeld, S. O., & Waldman, I. D. (Eds.). (2017). *Psychological science under scrutiny: Recent challenges and proposed solutions.* Hoboken, NJ: Wiley. doi:10.1002/9781119095910

Lindsay, D. S. (2015). Replication in psychological science. *Psychological Science, 26,* 1827-1832. doi:10.1177/0956797615616374

Loken, E., & Gelman, A. (2017). Measurement error and the replication crisis. *Science, 355*(6325), 584-585. doi: 10.1126/science.aal3618

Ly, A., Etz, A., Marsman, M., & Wagenmakers, E.-J. (2018). Replication Bayes factors from evidence updating. *Behavior Research Methods.* Advance online publication. doi:10.3758/s13428-018-1092-x

Makel, M. C., & Plucker, J. A. (2014). Facts are more important than novelty: Replication in the education sciences. *Educational Researcher, 43,* 304-316. doi:10.3102/0013189X14545513

Marín-Franch, I. (2018). Publication bias and the chase for statistical significance. *Journal of Optometry, 11,* 67-68. doi:10.1016/j.optom.2018.03.001

Martin, G. N., & Clarke, R. M. (2017). Are psychology journals anti-replication? A snapshot of editorial practices. *Frontiers in Psychology, 8,* 523. doi:10.3389/fpsyg.2017.00523

Martin, S. R., & Williams, D. R. (2017). Outgrowing the procrustean bed of normality: The utility of Bayesian modeling for asymmetrical data analysis. *PsyArXiv*, August 30. doi:10.31234/osf.io/26m49

Mathur, M. B., & VanderWeele, T. J. (2019). Challenges and suggestions for defining replication "success" when effects may be heterogeneous: Comment on Hedges and Schauer (2019). *Psychological Methods, 24,* 571-575. doi:10.1037/met0000223

Matthes, J., Marquart, F., Naderer, B., Arendt, F., Schmuck, D., & Adam, K. (2015). Questionable research practices in experimental communication research: A systematic analysis from 1980 to 2013. *Communication Methods and Measures, 9,* 193-207. doi:10.1080/19312458.2015.1096334

Maxwell, S. E. (2004). The persistence of underpowered studies in psychological research: Causes, consequences, and remedies. *Psychological Methods, 9,* 147-163. doi:10.1037/1082-989X.9.2.147

Maxwell, S. E., Lau, M. Y., & Howard, G. S. (2015). Is psychology suffering from a replication crisis? What does "failure to replicate" really mean? *American Psychologist, 70,* 487-498. doi:10.1037/a0039400

Meier, S. T. (1994). *The chronic crisis in psychological measurement and assessment: A historical survey.* San Diego, CA: Academic Press.

Merton, R. K. (1973). *The sociology of science, theoretical and empirical investigations.* Chicago, IL: University of Chicago Press.

Miller, D. J., & Hersen, M. (1992). *Research fraud in the behavioral and biomedical sciences.* New York, NY: Wiley.

Miller, R. G. (1981). *Simultaneous statistical inference* (2nd ed.). New York, NY: Springer.

Moore, D. S., McCabe, G. P., & Craig, B. A. (2017). *Introduction to the practice of statistics* (9th ed.). New York, NY: W. H. Freeman.

Morawski, J. (2019). The replication crisis: How might philosophy and theory of psychology be of use? *Journal of Theoretical and Philosophical Psychology, 39,* 218-238. doi:10.1037/teo0000129

Morey, R. D., Hoekstra, R., Rouder, J. N., Lee, M. D., & Wagenmakers, E.-J. (2016). The fallacy of placing confidence in confidence intervals. *Psychonomic Bulletin & Review, 23,* 103-123. doi:10.3758/s13423-015-0947-8

Munafò, M. R., Hollands, G. H., & Marteau, T. M. (2018). Open science prevents mindless science: Lessons from a case of academic misconduct. *BMJ, 363,* k4309. doi: 10.1136/bmj.k4309

Munafò, M. R., Nosek, B. A., Bishop, D. V. M., Button, K. S., Chambers, C. D., Percie du Sert, N., . . . Ioannidis, J. P. A. (2017). A manifesto for reproducible science. *Nature Human Behaviour, 1,* 21. doi:10.1038/s41562-016-0021

Murphy, K. R., & Aguinis, H. (2019). HARKing: How badly can cherry-picking and question trolling produce bias in published results? *Journal of Business and Psychology, 34,* 1-17. doi:10.1007/s10869-017-9524-7

National Academies of Sciences, Engineering, and Medicine (2019). *Reproducibility and replicability in science.* Washington, DC: The National Academies Press. doi:10.17226/25303

National Science Foundation (2018). *Science & Engineering Indicators 2018.* Online https://www.nsf.gov/statistics/2018/nsb20181

Nelson, L. D., Simmons, J., & Simonsohn, U. (2018). Psychology's Renaissance. *Annual Review of Psychology, 69,* 511-534. doi:10.1146/annurev-psych-122216-011836

Neuliep, J. W., & Crandall, R. (1990). Editorial bias against replication research. *Journal of Social Behavior and Personality, 5,* 85-90.

Neuliep, J. W., & Crandall R. (1993). Reviewer bias against replication research. *Journal of Social Behavior and Personality, 8,* 21-29.

Neyman, J., & Pearson, E. S. (1928). On the use and interpretation of certain test criteria for purposes of statistical inference. *Biometrika, 20A,* 175-240, 263-294. doi:10.2307/2331945

Neyman, J., & Pearson, E. S. (1933). On the problem of the most efficient tests of statistical hypotheses. *Transactions of the Royal Society of London Series A, 231,* 289-337. doi:10.1098/rsta.1933.0009

Norris, E., & O'Connor, D. B. (2019). Science as behaviour: Using a behaviour change approach to increase uptake of open science. *Psychology & Health, 34,* 1397-1406, doi:10.1080/08870446.2019.1679373

Nosek, B. A., Alter, G., Banks, G. C., Borsboom, D., Bowman, S. D., Breckler, S. J., ... Yarkoni, T. (2015). Promoting an open research culture. *Science, 348*(6242), 1422-1425. doi:10.1126/science.aab2374

Nosek, B. A., & Bar-Anan, Y. (2012). Scientific Utopia I: Opening scientific communication. *Psychological Inquiry, 23,* 217-243. doi:10.1080/1047840X.2012.692215

Nosek, B. A., Beck, E. D., Campbell, L., Flake, J. K., Hardwicke, T. E., Mellor, D. T., … Vazire, S. (2019). Preregistration is hard, and worthwhile. *Trends in Cognitive Sciences, 23,* 815-818. doi:10.1016/j.tics.2019.07.009

Nosek, B. A., Ebersole, C. R., DeHaven, A. C., & Mellor, D. T. (2018). The preregistration revolution. *Proceedings of the National Academy of Sciences of the United States of America, 115,* 2600-2606. doi: 10.1073/pnas.1708274114

Nosek, B. A., & Lakens, D. (2014). Registered reports: A method to increase the credibility of published results. *Social Psychology, 45,* 137-141. doi:10.1027/1864-9335/a000192

Nosek, B. A., Spies, J. R., & Motyl, M. (2012). Scientific utopia II: Restructuring incentives and practices to promote truth over publishability. *Perspectives on Psychological Science, 7,* 615-631. doi:10.1177/1745691612459058

Nuijten, M. B. (2019). Practical tools and strategies for researchers to increase replicability. *Developmental Medicine & Child Neurology, 61,* 535-539. doi:10.1111/dmcn.14054

Nuijten, M. B., Bakker, M., Maassen, E., & Wicherts, J. M. (2018). Verify original results through reanalysis before replicating. *Behavioral and Brain Sciences, 41,* e143. doi:10.1017/S0140525X18000791

Nuijten, M. B., Hartgerink, C. H. J., van Assen, M. A. L. M., Epskamp, S., & Wicherts, J. M. (2016). The prevalence of statistical reporting errors in psychology (1985-2013). *Behavior Research Methods, 48,* 1205-1226. doi:10.3758/s13428-015-0664-2

Oberauer, K., & Lewandowsky, S. (2019). Addressing the theory crisis in psychology. *Psychonomic Bulletin & Review, 26,* 1596-1618. doi: 10.3758/s13423-019-01645-2.

Onghena, P. (1998). *Methoden van onderzoek in de pedagogische wetenschappen: Empirisch-analytische methoden m.i.v. de statistiek, eerste deel.* Leuven: Acco.

Open Science Collaboration (2015). Estimating the reproducibility of psychological science. *Science, 349*(6251), aac4716.doi:10.1126/science.aac4716

Oreskes, N. (2019). *Why trust science?* Princeton, NJ: Princeton University Press.

Overbeke, A. J. P. M. (1994). Wangedrag in medisch-wetenschappelijk publiceren. *Nederlands Tijdschrift voor Geneeskunde, 138,* 1822-1826.

Pashler, H., & Harris, C. (2012). Is the replicability crisis overblown? Three arguments examined. *Perspectives on Psychological Science, 7,* 530-536. doi:10.1177/1745691612463401

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. doi: 10.1177/1745691612465253

Patil, P., Peng, R. D., & Leek, J. T. (2016). What should researchers expect when they replicate studies? A statistical view of replicability in psychological science. *Perspectives on Psychological Science, 11*, 539–544. doi: 10.1177/1745691616646366

Peels, R. (2019). Replicability and replication in the humanities. *Research Integrity and Peer Review, 4,* 2. doi:10.1186/s41073-018-0060-4

Phillips, C. V. (2004). Publication bias in situ. *BMC Medical Research Methodology, 4,* 20. doi:10.1186/1471-2288-4-20

Plesser, H. E. (2018). Reproducibility vs. replicability: A brief history of a confused terminology. *Frontiers in Neuroinformatics, 11,* 76. doi: 10.3389/fninf.2017.00076

Popper, K. R. (2007). *The logic of scientific discovery* [Logik der Forschung]. London, UK: Routledge. (Original work published 1935)

Pridemore, W. A., Makel, M. C., & Plucker, J. A. (2018). Replication in criminology and the social sciences. *Annual Review of Criminology, 1,* 19–38. doi:10.1146/annurev-criminol-032317-091849

Ranehill, E., Dreber, A., Johannesson, M., Leiberg, S., Sul, S., & Weber, R. A. (2015). Assessing the robustness of power posing: No effect on hormones and risk tolerance in a large sample of men and women. *Psychological Science, 26,* 653–656. doi:10.1177/0956797614553946

Ravitch, S. M., & Carl, N. M. (2016). *Qualitative research: Bridging the conceptual, theoretical and methodological.* Thousand Oaks, CA: SAGE Publications.

Reichenbach, H. (1938). *Experience and prediction: An analysis of the foundations and the structure of knowledge.* Chicago, IL: University of Chicago Press.

Reygel, P. (2019, 27 maart). Belangrijk dat de evolutieleer deel uitmaakt van het curriculum op school. *Knack*, online https://www.knack.be/nieuws/wetenschap/belangrijk-dat-de-evolutieleer-deel-uitmaakt-van-het-curriculum-op-school/article-opinion-1446119.html

Ritchie, S. J., Wiseman, R., & French, C. C. (2012). Failing the future: Three unsuccessful attempts to replicate Bem's 'retroactive facilitation of recall' effect. *PLoS ONE 7*(3), e33423. doi:10.1371/journal.pone.0033423

Rodgers, J. L., & Shrout, P. E. (2018). Psychology's replication crisis as scientific opportunity: A précis for policymakers. *Policy Insights from the Behavioral and Brain Sciences, 5,* 134-141. doi:10.1177/2372732217749254

Romero, F. (2018). Who should do replication labor? *Advances in Methods and Practices in Psychological Science, 1,* 516-537. doi:10.1177/2515245918803619

Rosenthal, R. (1979). The "file drawer problem" and tolerance for null results. *Psychological Bulletin, 86,* 638-641. doi:10.1037/0033-2909.86.3.638

Rossi, J. S. (1990). Statistical power of psychological research: What have we gained in 20 years? *Journal of Consulting and Clinical Psychology, 58,* 646-656. doi:10.1037//0022-006x.58.5.646

Rothstein, H. R., Sutton, A. J., & Borenstein, M. (Eds.) (2005). *Publication bias in meta-analysis: Prevention, assessment and adjustments.* Chichester, UK: Wiley.

Rubin, M. (2017). When does HARKing hurt? Identifying when different types of undisclosed post hoc hypothesizing harm scientific progress. *Review of General Psychology, 21,* 308-320. doi:10.1037/gpr0000128

Sanborn, A. N., & Hills, T. T. (2014). The frequentist implications of optional stopping on Bayesian hypothesis tests. *Psychonomic Bulletin & Review, 21,* 283-300. doi:10.3758/s13423-013-0518-9

Savin-Baden, M., & Major, C. H. (2013). *Qualitative research : the essential guide to theory and practice.* New York, NY: Routledge.

Scheufele, D. A. (2014). Science communication as political communication. *Proceedings of the National Academy of Sciences of the United States of America, 111*(Suppl. 4), 13585-13592. doi:10.1073/pnas.1317516111

Scheufele, D. A., & Krause, N. M. (2019). Science audiences, misinformation, and fake news. *Proceedings of the National Academy of Sciences of the United States of America, 116*, 7662-7669. doi:10.1073/pnas.1805871115

Schimmack, U. (2019). *The validation crisis in psychology.* Unpublished manuscript https://replicationindex.files.wordpress.com/2019/04/validation.crisis.v3.pdf

Schloss, P. D. (2018). Identifying and overcoming threats to reproducibility, replicability, robustness, and generalizability in microbiome research. *mBio, 9*(3):e00525-18. doi:10.1128/mBio.00525-18

Schmidt, B., Ross-Hellauer, T., van Edig, X., & Moylan, E. C. (2018). Ten considerations for open peer review. *F1000Research, 7,* 969. doi:10.12688/f1000research.15334.1

Schmidt, S. (2009). Shall we really do it again? The powerful concept of replication is neglected in the social sciences. *Review of General Psychology, 13,* 90-100. doi:10.1037/a0015108

Schonbrodt, F. D., Wagenmakers, E.-J., Zehetleitner, M., & Perugini, M. (2017). Sequential hypothesis testing with Bayes factors: Efficiently testing mean differences. *Psychological Methods, 22,* 322-339. doi:10.1037/met0000061

Schott, E., Rhemtulla, M., & Byers-Heinlein, K. (2019). Should I test more babies? Solutions for transparent data peeking. *Infant Behavior & Development, 54,* 166-176. doi:10.1016/j.infbeh.2018.09.010

Schulson, M. (2018). Science's "reproducibility crisis" is being used as political ammunition. *Wired,* online https://www.wired.com/story/sciences-reproducibility-crisis-is-being-used-as-political-ammunition/

Sedlmeier, P., & Gigerenzer, G. (1989). Do studies of statistical power have an effect on the power of studies? *Psychological Bulletin, 105,* 309-316. doi:10.1037/0033-2909.105.2.309

Segers, J. H. G. (1989). *Methoden voor de sociale wetenschappen, deel 1* (5de druk). Assen, NL: Van Gorcum.

Seglen, P. O. (1997). Why the impact factor of journals should not be used for evaluating research. *BMJ, 314,* 498-502. doi:10.1136/bmj.314.7079.497

Shadish, Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference.* Houghton, MI: Mifflin and Company.

Shapin, S., & Schaffer, S. (1985). *Leviathan and the air-pump: Hobbes, Boyle, and the experimental life.* Princeton, NJ: Princeton University Press.

Shrout, P. E., & Rodgers, J. L. (2018). Psychology, science, and knowledge construction: Broadening perspectives from the replication crisis. *Annual Review of Psychology, 69,* 487-510. doi:10.1146/annurev-psych-122216-011845

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. doi:10.1177/0956797611417632

Simmons, J. P., & Simonsohn, U. (2017). Power posing: P-curving the evidence. *Psychological Science, 28,* 687-693. doi:10.1177/0956797616658563

Simons, D. J., Shoda, Y., & Lindsay, D. S. (2017). Constraints on Generality (COG): A proposed addition to all empirical papers. *Perspectives on Psychological Science, 12,* 1123-1128. doi:10.1177/1745691617708630

Simonsohn, U. (2015). Small telescopes: Detectability and the evaluation of replication results. *Psychological Science, 26*, 559-569. doi:10.1177/0956797614567341

Skinner, G. (2019, September 18). Trust: The truth. *Ipsos MORI*, online https://www.ipsos.com/ipsos-mori/en-uk/ipsos-thinks-trust-truth

Slow Science Academy (2010). The Slow Science Manifesto. Slow Science Academy, online http://slow-science.org/

Smaldino, P. E., & McElreath, R. (2016). The natural selection of bad science. *Royal Society Open Science, 3,* 160384. doi:10.1098/rsos.160384

Smith, N. C. (1970). Replication studies: a neglected aspect of psychological research. *American Psychologist, 25,* 970-975. doi: 10.1037/h0029774

Soares, M. J., Müller, M. J., Boeing, H., Maffeis, C., Misra, A., Muscogiuri, G., …, Zhu S. (2019). Conflict of interest in nutrition research: an editorial perspective. *European Journal of Clinical Nutrition, 73,* 1213-1215. doi:10.1038/s41430-019-0488-8

Soto, C. J. (2019). How replicable are links between personality traits and consequential life outcomes? The Life Outcomes of Personality Replication Project. *Psychological Science, 30,* 711-727. doi:10.1177/0956797619831612

Srivastava, S. (2018). Verifiability is a core principle of science. *Behavioral and Brain Sciences, 41,* e150. doi: 10.1017/S0140525X18000869

Stanley, T. D., Carter, E. C., & Doucouliagos, H. (2018). What meta-analyses reveal about the replicability of psychological research. *Psychological Bulletin, 144,* 1325-1346. doi:10.1037/bul0000169

Steijaert, M. (2017, 24 juni). Bij deze voedingsprofessor komen alle problemen van de wetenschap samen. *De Volkskrant*, online https://www.volkskrant.nl/wetenschap/bij-deze-voedingsprofessor-komen-alle-problemen-van-de-wetenschap-samen~bf542fa4/

Steijaert, M. (2018, 21 september). Omstreden voedingsprof Brian Wansink stapt op na intern onderzoek. *De Volkskrant*, online https://www.volkskrant.nl/wetenschap/omstreden-voedingsprof-brian-wansink-stapt-op-na-intern-onderzoek~bdcd1162/

Sterling, T. D. (1959). Publication decisions and their possible effects on inferences drawn from tests of significance – or vice versa. *Journal of the American Statistical Association 54,* 30–34. doi:10.2307/2282137

Sterling, T. D., Rosenbaum, W. L. & Weinkam, J. J. (1995). Publication decisions revisited: The effect of the outcome of statistical tests on the decision to publish and vice versa. *The American Statistician, 49,* 108-112. doi:10.1080/00031305.1995.10476125

Stroebe, W., & Strack, F. (2014). The alleged crisis and the illusion of exact replication. *Perspectives on Psychological Science, 9,* 59-71. doi:10.1177/1745691613514450

Sumner, P., Vivian-Griffiths, S., Boivin, J., Williams, A., Bott, L., Adams, R, …, Chambers, C. D. (2016). Exaggerations and caveats in press releases and health-related science news. PLoS ONE, 11(12), e0168217. doi:10.1371/journal.pone.0168217

Tackett, J. L., & McShane, B. B. (2018). Conceptualizing and evaluating replication across domains of behavioral research. *Behavioral and Brain Sciences, 41,* e152. doi: 10.1017/S0140525X18000882

ten Hagen, S. L. (2019). How "facts" shaped modern disciplines: The fluid concept of fact and the common origins of German physics and historiography. *Historical Studies in the Natural Sciences, 49,* 300-337. doi:10.1525/hsns.2019.49.3.300

The Academy of Medical Sciences (2015). *Reproducibility and reliability of biomedical research: improving research practice.* London, UK: The Academy of Medical Sciences.

The PLoS Medicine Editors (2006) The impact factor game. *PLoS Medicine, 3*(6), e291. doi:10.1371/journal.pmed.0030291

Tressoldi, P. (2012). Replication unreliability in psychology: Elusive phenomena or "elusive" statistical power? *Frontiers in Psychology, 3,* 218. doi:10.3389/fpsyg.2012.00218

Tressoldi, P. E., & Giofré, D. (2015). The pervasive avoidance of prospective statistical power: Major consequences and practical solutions. *Frontiers in Psychology, 6,* 726. doi:10.3389/fpsyg.2015.00726

Uhlmann, E. L., Ebersole, C. R., Chartier, C. R., Errington, T. M., Kidwell, M. C., Lai, C. K., … Nosek, B. A. (2019). Scientific Utopia III: Crowdsourcing science. *Perspectives on Psychological Science, 14,* 711-733. doi:10.1177/1745691619850561

Van Bavel, J. J., Mende-Siedlecki, P., Brady, W. J. & Reinero, D. A. (2016). Contextual sensitivity in scientific reproducibility. *Proceedings of the National Academy of Sciences of the United States of America, 113,* 6454-6459. doi:10.1073/pnas.1521897113

van der Zee, T., Anaya, J., & Brown, N. J. L. (2017). Statistical heartburn: An attempt to digest four pizza publications from the Cornell Food and Brand Lab. *BMC Nutrition, 3,* 54. doi:10.1186/s40795-017-0167-x

van der Zee, T., & Reich, J. (2018). Open Education Science. *AERA Open, 4*(3), 1-15. doi:10.1177/2332858418787466

Vankov, I., Bowers, J., & Munafò, M. R. (2014). On the persistence of low power in psychological science. *The Quarterly Journal of Experimental Psychology, 67,* 1037-1040. doi:10.1080/17470218.2014.885986

Van Liedekerke, A., Van Driessche, V., & Nollet, V. (2019). Voorschriften wetenschappelijke integriteit nog weinig afgedwongen: Fraudebestrijding klimt wel hoger op agenda. *Veto,* 20 mei 2019, http://www.veto.be/artikel/voorschriften-wetenschappelijke-integriteit-nog-weinig-afgedwongen

Vazire, S. (2018). Implications of the credibility revolution for productivity, creativity, and progress. *Perspectives on Psychological Science, 13,* 411-417. doi:10.1177/1745691617751884

Vul, E., Harris, C., Winkielman, P., & Pashler, H. (2009). Puzzlingly high correlations in fMRI studies of emotion, personality, and social cognition. *Perspectives on Psychological Science, 4,* 274-290. doi:10.1111/j.1745-6924.2009.01125.x

Wagenmakers, E.-J. (2007). A practical solution to the pervasive problems of *p* values. *Psychonomic Bulletin & Review, 14,* 779-804. doi:10.3758/bf03194105

Wagenmakers, E.-J., Wetzels, R., Borsboom, D., van der Maas, H. L. J., & Kievit, R. A. (2012). An agenda for purely confirmatory research. *Perspectives on Psychological Science, 7,* 632-638. doi:10.1177/1745691612463078

Wald, A. (1947). *Sequential analysis.* New York, NY: Wiley.

Wansink, B. (2006). *Mindless eating: Why we eat more than we think.* New York, NY: Bantam.

Warren, M. (2018, October 24). First analysis of 'pre-registered' studies shows sharp rise in null findings. *Nature News*, online https://www.nature.com/articles/d41586-018-07118-1 doi:10.1038/d41586-018-07118-1

Westfall, P. H., & Young, S. S. (1993). *Resampling-based multiple testing: Examples and methods for* p-*value adjustment.* New York, NY: Wiley.

Wicherts, J. M. (2011). Psychology must learn a lesson from fraud case. *Nature, 480,* 7. doi:10.1038/480007a

Wicherts, J. M. (2017). The weak spots in contemporary science (and how to fix them). *Animals, 7,* 90. doi:10.3390/ani7120090

Wicherts, J. M., Veldkamp, C. L., Augusteijn, H. E., Bakker, M., van Aert, R. C., & van Assen, M. A. (2016). Degrees of freedom in planning, running, analyzing, and reporting psychological studies: A checklist to avoid p-hacking. *Frontiers in Psychology, 7,* 1832. doi:10.3389/fpsyg.2016.01832

Wiersma, W. (1995). *Research methods in education: An introduction* (6th ed.). Boston, MA: Allyson & Bacon.

Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., …, Mons, B. (2016). The FAIR guiding principles for scientific data management and stewardship. *Scientific Data, 3,* 160018. doi:10.1038/sdata.2016.18

Wingen, T., Berkessel J. B., & Englich, B. (2019). No replication, no trust? How low replicability influences trust in psychology. *Social Psychological and Personality Science,* online preprint doi:10.1177/1948550619877412

Wiseman, R., Watt, C., & Kornbrot, D. (2019). Registered reports: An early example and analysis. *PeerJ, 7,* e6232. doi: 10-.7717/peerj.6232

Zwaan, R. A., Etz, A., Lucas, R. E., & Donnellan, M. B. (2018). Making replication mainstream. *Behavioral and Brain Sciences, 41,* e120. doi: 10.1017/S0140525X17001972

RECENTE STANDPUNTEN (vanaf 2016)

42. Erik Mathijs, Willy Verstraete (e.a.), *Vlaanderen wijs met water: waterbeleid in transitie*, KVAB/ Klasse Technische wetenschappen, 2016.

43. Erik Schokkaert - *De gezondheidszorg in evolutie: uitdagingen en keuzes*, KVAB/Klasse Menswetenschappen, 2016.

44. Ronnie Belmans, Pieter Vingerhoets, Ivo Van Vaerenbergh e.a. – *De eindgebruiker centraal in de energietransitie*, KVAB/Klasse Technische Wetenschappen, 2016.

45. Willem Elias, Tom De Mette – *Doctoraat in de kunsten*, KVAB/Klasse Kunsten, 2016.

46. Hendrik Van Brussel, Joris De Schutter e.a., *Naar een inclusieve robotsamenleving*, KVAB/Klasse Technische Wetenschappen, 2016.

47. Bart Verschaffel, Marc Ruyters e.a., *Elementen van een duurzaam kunstenbeleid*, KVAB/Klasse Kunsten, 2016.

48. Pascal Verdonck, Marc Van Hulle (e.a.) - *Datawetenschappen en gezondheidszorg*, KVAB/Klasse Technische wetenschappen, 2017.

49. Yolande Berbers, Mireille Hildebrandt, Joos Vandewalle (e.a.) - *Privacy in tijden van internet, sociale netwerken en big data*, KVAB/Klasse Technische wetenschappen, 2017.

50. Barbara Baert (e.a.), *Iconologie of 'La science sans nom'*, KVAB/Klasse Kunsten, 2017.

51. Tariq Modood, Frank Bovenkerk – *Multiculturalism. How can Society deal with it?* KVAB/Klasse Menswetenschappen, 2017.

52. Mark Eyskens – *Europa in de problemen.* KVAB/Klasse Menswetenschappen, 2017.

53. Luc Steels – *Artificiële intelligentie. Naar een vierde industriële revolutie?.* KVAB/Klasse Natuurwetenschappen, 2017.

54. Godelieve Gheysen, René Custers, Dominique Van Der Straeten, Dirk Inzé*, Ggo's anno 2018. Tijd voor een grondige herziening*. KVAB/Klasse Natuurwetenschappen, 2017.

55. Christoffel Waelkens (e.a.) – *Deelname van Vlaanderen aan grote internationale onderzoeksinfrastructuren: uitdagingen en aanbevelingen*, KVAB/Klasse Natuurwetenschappen, 2017.

55. Addendum. Jean-Pierre Henriet. – *Mijlpalen in internationale wetenschappelijke samenwerking*, KVAB/ Klassen Natuurwetenschappen, 2017.

56. Piet Swerts, Piet Chielens, Lucien Posman – *A Symphony of Trees. Wereldcreatie naar aanleiding van de herdenking van de Derde Slag bij Ieper, 1917*, KVAB/Klasse Kunsten, 2017.

57. Willy Van Overschée e.a. – *De mobiliteit van morgen: zijn we klaar voor een paradigmawissel*?, KVAB/Klasse Technische Wetenschappen, 2018.

59. Dirk Van Dyck, Elisabeth Monard, Sylvia Wenmackers e.a. – *Onderzoeker-gedreven wetenschap. Analyse van de situatie in Vlaanderen*, KVAB/Klasse Natuurwetenschappen, 2018.

60. Liliane Schoofs – *Doctoraathouders geven het Vlaanderen van morgen vorm*, KVAB/Klasse Natuurwetenschappen, 2018.

61. Luc Bonte, Aimé Heene, Paul Verstraeten e.a. – *Verantwoordelijk omgaan met digitalisering. Een oproep naar overheden en bedrijfsleven, waar ook de burger toe kan/moet bijdragen*, KVAB/Klasse Technische Wetenschappen, 2018.

62. Jaak Billiet, Michaël Opgenhaffen, Bart Pattyn, Peter Van Aelst – *De strijd om de waarheid. Over nepnieuws en desinformatie in de digitale mediawereld*, KVAB/Klasse Menswetenschappen, 2018.

63. Christoffels Waelkens. – *De Vlaamse Wetenschapsagenda en interdisciplinariteit. Leren leven met interdisciplinaire problemen en oplossingen*, KVAB/Klasse Natuurwetenschappen, 2020.

65. Mark Eyskens – *Als een virus de mensheid gijzelt. Oorzaken en gevolgen van de Coronacrisis*, KVAB/ Klasse Menswetenschappen, 2020.

Is er naast fake news ook zoiets als fake science? Van wetenschappelijke bevindingen verwachten we dat ze repliceerbaar zijn. Maar in de empirische menswetenschappen blijken deze niet altijd stand te houden in onafhankelijk vervolgonderzoek. Bovendien blijkt dat wanneer er toch een effect wordt gevonden, dit effect meestal kleiner is dan in de oorspronkelijke studie.

Dit standpunt gaat dieper in op de uitgangspunten en het begrippenkader waarin deze zogenaamde “repliceerbaarheidscrisis” gesitueerd is. We onderzoeken de antecedenten en de mogelijke oorzaken van de beperkte repliceerbaarheid en bespreken de implicaties voor de geloofwaardigheid van de empirische menswetenschappen. Op basis van die analyse wordt getracht om een weg uit de crisis te vinden. Zoals psychologen beweren: elke crisis bevat de kiem voor groei. We hopen dat deze groei repliceerbaar is.

De reeks Standpunten van de Academie is een bijdrage tot het wetenschappelijk onderbouwd debat over actuele maatschappelijke en artistieke thema’s. De auteurs, leden en werkgroepen van de Academie schrijven in eigen naam, onafhankelijk en met volledige intellectuele vrijheid. De goedkeuring voor publicatie door een of meerdere Klassen van de Academie waarborgt de kwaliteit van de gepubliceerde studies.

Vlaamse overheid

RICHTING MORGEN
MET WETENSCHAP, TECHNOLOGIE, CREATIVITEIT & INNOVATIE