rijksuniversiteit groningen

Vertrouwen in wetenschappelijk onderzoek

Een significante voorkeur

Student: S.A. Daalder (S3233383)

Begeleider: dr. R. Hoekstra, A.H. Stoevenbelt MSc.
2e beoordelaar: dr. J. Luijkx

Rijksuniversiteit Groningen
Faculteit der Gedrags- en Maatschappijwetenschappen
Bachelorwerkstuk Pedagogische Wetenschappen
Juni 2022

Abstract

Publication bias appears to have a negative impact on scientific research, as studies with positive significant outcomes are more likely to be published regardless of quality compared to negative non-significant outcomes (Greenwald, 1975; Moss & De Bin, 2021; Rosenthal, 1979; Torgerson, 2006). Our students are the scientists of the future. It is important to uncover the preferences of students for significant or non-significant p-values, in order to be able to assess whether the trend in the scientific literature also affects their assessment of articles. The following question was asked: “To what extent do students of the Faculty of Behavioral and Social Sciences (University of Groningen) find the p-value important for rating the quality of scientific articles?”

To answer the research question, a quantitative study was conducted. Ten scientific articles, with five significant and five non-significant outcomes, were presented to the respondents using a closed survey by Qualtrics. The results showed that a significant effect ($t(33)=1.783$, $p = 0.042$) was probably found among students of the Faculty of Behavioral and Social Sciences during the rating of the p-value for the quality of scientific articles, but the effect size (.31) between significant and non-significant p-values turned out to be minimal.

A limitation within this study concerns the operationalization of the word 'important'. It is therefore recommended to better operationalize the word 'important'. Secondly, it is recommended to adjust the included p-values to study the Cliff effect. Thirdly, more respondents could be recruited, so that the statistical power of this study could be strengthened. Finally, no distinction was made within the different studies of the Faculty of Behavioral and Social Sciences. To conclude, there was a minimal preference for significance.

Inleiding

Wanneer iemand onderzoek doet, moet deze diens beweringen onderbouwen met bewijs. Ook al worden de uitkomsten van wetenschappelijk onderzoek vaak als stellig gepresenteerd in de literatuur, is die stelligheid niet altijd verantwoord, omdat er sprake kan zijn van publicatiebias. Waar publicatiebias optreedt, is het bewijs voor een bepaalde bewering niet zo sterk als we zouden willen, omdat sommige resultaten in de praktijk niet worden ondersteund door de literatuur. De literatuur, met veel significante uitkomsten, is niet altijd een afspiegeling van wat er daadwerkelijk binnen een wetenschappelijk onderzoek gevonden worden. In de praktijk worden namelijk zowel significante als niet-significante resultaten gevonden. Er blijkt een sterke relatie te zijn tussen de resultaten van een onderzoek en of het onderzoek is gepubliceerd, deze relatie kan wijzen op publicatiebias (Franco, Malhotra & Simonovits, 2014).

Publicatiebias is het fenomeen binnen wetenschappelijk onderzoek, waarbij er de tendens is om een groot deel statistisch significante positieve resultaten (met een kleine p-waarde) te publiceren ongeacht de kwaliteit van onderzoek (Rosenthal, 1979). Waardoor, omgekeerd, een groter deel statistisch significante negatieve of nulresultaten (met een hogere p-waarde) niet wordt gepubliceerd (Greenwald, 1975; Moss & De Bin, 2021; Rosenthal, 1979; Torgerson, 2006). Het zorgt voor een vertekening binnen wetenschappelijk onderzoek (Greenwald, 1975; Moss & De Bin, 2021; Rosenthal, 1979; Torgerson, 2006). Demaria (2004) voegt aan deze definitie toe dat ook publicaties die worden vertraagd, die verschijnen in niet-geïndexeerde tijdschriften of die genegeerd worden door de media onder publicatiebias kunnen vallen. In het voorbeeld dat Ferguson & Brannick (2012) schetsen in hun artikel, waarbij er in de casus sprake was van selectieve publicatie over de werkzaamheid van antidepressiva, dat er meer nadruk is gekomen op de verstorende effecten die publicatiebias heeft op wetenschappelijke artikelen.

Publicatiebias is een probleem voor de geloofwaardigheid van wetenschappelijk onderzoek en de literatuur. Het is namelijk één van de grootste bedreigingen voor de validiteit van resultaten, omdat publicatiebias kan leiden tot overwaardering van effecten en kan zorgen voor type I-fouten (vals-positief) (Franco, Malhotra & Simonovits, 2014; Ioannidis, 2005; Van Aert, Wicherts, van Assen & Macleod, 2019). Dit houdt in dat een experiment ten onrechte positief of significant wordt beoordeeld, terwijl de uitkomsten niet overeenkomen met de werkelijkheid. Onderzoeken met type I-fouten worden volgens Rosenthal (1979) eerder gepubliceerd dan onderzoeken met niet-significante uitkomsten. Door deze negatieve studies echter wel te publiceren, kan er kritisch gekeken worden naar problemen of uitkomsten die door

anderen zijn aangesneden (Miller & Moulder, 1998). Dus, negatieve studies kunnen ook nuttig zijn om uiteindelijk het 'niveau van bewijs' te verhogen (Miller & Moulder, 1998).

Het blijkt dat gepubliceerde studies in de sociale en gedragswetenschappen vaker met ondersteunende (significante) hypothesen worden gepubliceerd dan in de fysische of biologische wetenschappen (Fanelli, 2010). Een reden waarom er sprake kan zijn van publicatiebias, is omdat 'positieve studies' (die statistisch significant zijn) interessanter voor onderzoekers zijn dan 'negatieve studies' (die niet statistisch significant zijn), aangezien het vaak leidt tot openbaringen en een mogelijkheid tot verder onderzoek (Rockwell, Kimler & Moulder, 2006). Uit het onderzoek van Kicinski (2013) komt naar voren dat studies die statistisch significante uitkomsten aantoonden vaak een hogere kans hadden om in recente meta-analyses geplaatst te worden, dan studies die andere uitkomsten lieten zien. Onderzoeken met statistisch significante uitkomsten hebben niet alleen meer kans om gepubliceerd te worden, maar deze onderzoeken worden ook vaker geciteerd en gepromoot (Marin-Franch, 2018). Het blijkt dat onderzoekers dus meer vertrouwen hebben in significante uitkomsten en deze liever gebruiken voor eigen onderzoek. In de literatuur wordt ook wel gesproken over een Cliff-effect als er sprake is van een daling in vertrouwen als de p-waarde voorbij de kritieke waarde ($p > 0.05$) komt (Nelson, Rosenthal & Rosnow, 1986).

Studenten in het hoger onderwijs zijn de wetenschappers van de toekomst. Deze studenten zijn onderdeel van én maken vaak gebruik van wetenschappelijk onderzoek. Studenten zijn meer geneigd literatuur te gebruiken, als zij deze als relevant beschouwen en vertrouwen hebben in het proces dat tot de resultaten is gekomen (Mullins, Crowe & Wymer, 2010). Een onderdeel van het proces dat uiteindelijk het vertrouwen in de wetenschap van zowel studenten als onderzoekers kan beïnvloeden, is volgens Mullins, Crowe & Wymer (2010) de problemen met betrekking tot publicatiebias. In een eerder onderzoek van Rosenthal & Gaito (1963) bleek dat de meerderheid van studenten bij $p > 0.05$ kenmerken vertoonde van een Cliff-effect, door minder vertrouwen uit te spreken over het significantieniveau. De impact van publicatiebias wordt al meerdere decennia als belangrijk beschouwd (Rosenthal, 1979). Hoe studenten, die vrijwel dezelfde studieachtergrond hebben (gedrags- en maatschappijwetenschappen), verschillen in de mate van voorkeur betreffende publicatiebias, is dan ook van maatschappelijke relevantie, omdat zij of doorgaan binnen de wetenschap of wetenschappelijk onderzoek gebruiken in hun werkveld.

De hoofdvraag van dit bachelorwerkstuk luidt daarom als volgt: "In hoeverre vinden studenten van de faculteit Gedrags- en Maatschappijwetenschappen (Rijksuniversiteit

Groningen) de p-waarde belangrijk voor het beoordelen van de kwaliteit van wetenschappelijke artikelen?"

Er wordt verwacht dat studenten van de faculteit Gedrags- en Maatschappijwetenschappen wetenschappelijke artikelen met een significante p-waarde als belangrijker beoordelen dan wetenschappelijke artikelen met een niet-significante p-waarde, aangezien deze studenten wetenschappelijk gefocust zijn. De verwachting is dat deze studenten de p-waarde belangrijk achten voor de kwaliteit en hiermee hun vertrouwen uitspreken over wetenschappelijk onderzoek. Binnen het onderzoek zou er mogelijk sprake kunnen zijn van een Cliff-effect, waarbij studenten de p-waarde minder vertrouwen als deze voorbij de kritieke waarde van $p > 0.05$ komt.

Methode

Om de onderzoeksvraag: "In hoeverre vinden studenten van de faculteit Gedrags- en Maatschappijwetenschappen (Rijksuniversiteit Groningen) de p-waarde belangrijk voor het beoordelen van de kwaliteit van wetenschappelijke artikelen?" te beantwoorden, is een kwantitatief onderzoek uitgevoerd.

Respondenten

De onderzoeks- en doelpopulatie waren studenten die op het moment van onderzoek studeerden aan de faculteit Gedrags- en Maatschappijwetenschappen aan de Rijksuniversiteit Groningen. Om deze studenten te werven is er gebruik gemaakt van een convenience steekproef. Er werd getracht om minimaal een steekproefgrootte van 100 respondenten te verkrijgen. Uiteindelijk hebben er 34 respondenten geparticipeerd aan het onderzoek, door de vragenlijst te beantwoorden. Alle respondenten ($N = 81$) die de vragenlijst hebben opengelaten of niet volledig hebben afgerond zijn geëxcludeerd uit dit onderzoek. De verwachting is dat de respondenten voornamelijk studenten van Pedagogische Wetenschappen en de Academische Opleiding Leraar Basisonderwijs geweest zullen zijn, aangezien deze behoren tot het eigen netwerk. Dit en andere demografische kenmerken zijn echter niet gevraagd aan de respondenten, in verband met de privacy van de respondenten.

Procedure

Voor dit onderzoek werd gebruik gemaakt van een vragenlijst om een antwoord te vinden op de onderzoeksvraag door het verzamelen van nieuwe data. Er is een gesloten enquête verspreid onder studenten van de faculteit Gedrags- en Maatschappijweten via Qualtrics. De enquête is verspreid via het eigen netwerk, door middel van het direct benaderen van studiegenoten en door het verspreiden via sociale media (Facebook, Instagram en LinkedIn). Respondenten hebben tot 20 mei 2022 de tijd gehad om de enquête in te vullen. Het invullen

van de enquête duurde ongeveer 15 minuten. Alleen volledig ingevulde enquêtes, en die volgens de informed consent toestemming hebben gegeven, zijn meegenomen in dit onderzoek. Studenten werden van tevoren niet ingelicht over het doel van het onderzoek, zodat het onderzoek niet werd beïnvloed met sociaal wenselijke antwoorden.

Om te voldoen aan de Algemene Verordening Gegevensbescherming (AVG) zijn de gegevens anoniem verzameld en opgeslagen op de beveiligde schijf van de Rijksuniversiteit Groningen.

Materialen

In de gesloten enquête werd eerst informatie gegeven over het onderzoek, waarna tien artikelen, met vijf significante en vijf niet-significante wetenschappelijke artikelen werden getoond (zie Bijlage). De respondenten hebben hierbij vragen gekregen over wat zij, op een vijfpunts Likert-schaal (1 = zeer onbelangrijk, 2 = belangrijk, 3 = neutraal, 4 = belangrijk, 5 = zeer belangrijk) vinden van de steekproefgrootte, de betrouwbaarheid van de gebruikte schaal aan de hand van Cronbachs alfa, de p-waarde en het uitgavejaar van het artikel. Tot slot heeft er een debriefing plaats gevonden, waarin onder andere werd vermeld dat sommige wetenschappelijke artikelen (deels) aangepaste gegevens bevatten. Voor dit onderzoek was alleen de vraag over de p-waarde interessant. De andere vragen werden gebruikt als afleiders, zodat de respondenten niet gelijk in de gaten zouden hebben wat het precieze onderzoeksdoel was. Daarnaast is er sprake geweest van counterbalancing van de artikelen en de genoemde p-waarden, zodat de scores die een respondent gaf aan de p-waarde niet afhankelijk waren van de tekst van een artikel dat erbij hoorde. De respondenten zijn verdeeld over de tien verschillende blocken. De vijfpunts Likert-schaal had een geschatte betrouwbaarheid van Guttman's lambda-2 = 0.914.

De inhoudsvaliditeit is binnen dit onderzoek gewaarborgd, aangezien de p-waarde de mate van significantie aangeeft en omdat de tien artikelen in de enquête zijn gebaseerd op daadwerkelijke artikelen die geschikt geacht werden voor de onderzoeks- en doelpopulatie.

Data-analyse

Bij de data-analyse zijn de data samengevat met behulp van beschrijvende statistiek (zie Tabel 1). Er is bekeken wat de minimale, maximale, gemiddelde score en standaarddeviatie per p-waarde waren. Vervolgens zijn er gemiddelde scores berekend voor de vijf significante p-waardes en de vijf niet-significantie p-waardes, eveneens is de beschrijvende statistiek van deze gemiddelde scores waargenomen. Ook is er gekeken naar de beschrijvende statistiek voor de totaalscore van alle p-waarden. Daarna werd een eenzijdige matched pairs t-test uitgevoerd. Er is gekozen voor een eenzijdige matched pairs t-test aangezien de hypothese verondersteld

dat er zich een verschil zou voordoen, waarbij wetenschappelijke artikelen met significante p-waardes belangrijker werden geacht dan wetenschappelijke artikelen met niet-significante p-waardes. Voor de matched pairs t-test mag ervan worden uitgegaan dat de assumptie van onafhankelijke respondenten is voldaan. Bovendien is voldaan aan de assumptie dat er sprake is geweest van een normale verdeling van de gemiddelde scores (zie Figuur 2). Met deze analyse is bekeken of de scores van respondenten verschillen op de beoordeling van de significante en niet-significante p-waardes. In de hypothese werd verondersteld dat er een verschil zou voordoen, waarbij wetenschappelijke artikelen met significante p-waardes als belangrijker zouden worden beoordeeld dan wetenschappelijke artikelen met niet-significante p-waardes.

Resultaten

Naar aanleiding van de onderzoeksvraag: “In hoeverre vinden studenten van de faculteit Gedrags- en Maatschappijwetenschappen (Rijksuniversiteit Groningen) de p-waarde belangrijk voor het beoordelen van de kwaliteit van wetenschappelijke artikelen?” zijn er tien artikelen met verschillende p-waardes voorgelegd aan studenten van de faculteit Gedrags- en Maatschappijwetenschappen. In Tabel 1 is af te lezen wat de minimale score, maximale score, gemiddelde score en standaarddeviatie per p-waarde, per subgroep (significant of niet-significant) en in totaal waren. Gedurende het onderzoek naar de beoordeling van de verschillende p-waardes geldt dat studenten niet de waarde 1.00 (zeer onbelangrijk) hebben toegekend aan een p-waarde. Als er naar de individuele p-waardes gekeken wordt, valt hierin op dat de kleinste p-waardes (p = 0.001 en p = 0.002), gemiddeld hoger (4.12 en 4.06) zijn beoordeeld dan de andere p-waardes. Ook is af te lezen dat het gemiddelde van de significante p-waardes (3.89) 0,03x hoger beoordeeld werd ten opzichte van het gemiddelde van niet-significante waardes (3.76).

Tabel 1

Centrum- en spreidingsmaten beoordelingen p-waardes

	Minimum	Maximum	Gemiddelde	Standaarddeviatie
$p = 0.03$	2.00	5.00	3.76	0.92
$p = 0.23$	2.00	5.00	3.82	0.97
$p = 0.54$	2.00	5.00	3.74	1.05
$p = 0.002$	2.00	5.00	4.06	0.95
$p = 0.02$	2.00	5.00	3.76	0.96
$p = 0.38$	2.00	5.00	3.74	0.86
$p = 0.069$	2.00	5.00	3.71	0.72
$p = 0.166$	2.00	5.00	3.79	0.88
$p = 0.049$	2.00	5.00	3.74	0.90
$p = 0.001$	2.00	5.00	4.12	0.95
Significante p-waardes	2.00	4.80	3.89	0.75
Niet-significant p-waardes	2.00	4.80	3.76	0.68
Totaalscore p-waardes	2.00	4.70	3.82	0.68

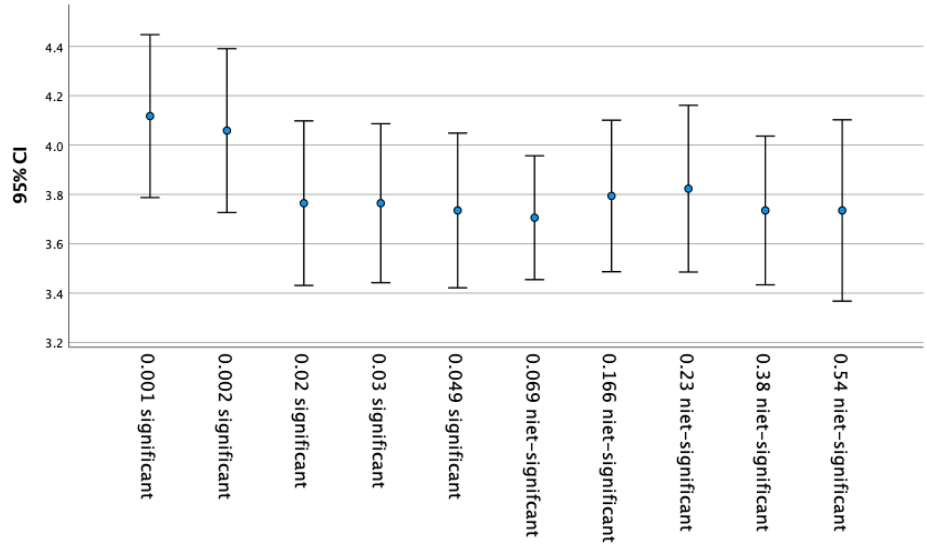
Voetnoot: $N = 34$

Cliff-effect

Om te kunnen beoordelen of er een Cliff-effect in de beoordeling van de p-waardes heeft voorgedaan, is in Figuur 1 een plot zichtbaar gemaakt. Uit de literatuur is gebleken dat een Cliff-effect, een plotselinge daling in vertrouwen, over het algemeen zichtbaar is bij de kritieke waarde ($p > 0.05$) (Nelson, Rosenthal & Rosnow, 1986). In Figuur 1 valt te zien dat er niet een extreem Cliff-effect zich heeft voorgedaan, maar dat de beoordelingen vrijwel op één lijn liggen. Wel is zichtbaar dat $p = 0.001$ en $p = 0.002$ aanmerkelijk hoger liggen.

Figuur 1

Beoordeling p-waardes

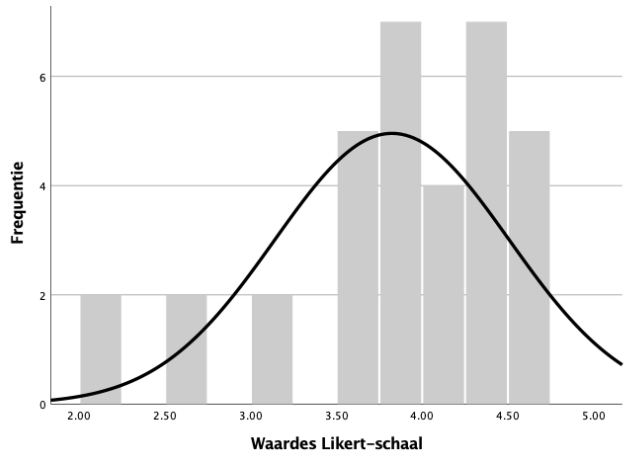


Toetsen hypothesen

Om de hypothesen behorend bij de onderzoeksvraag te toetsen, is er gebruik gemaakt van een eenzijdige matched pairs t-test. Bij dit onderzoek is voldaan aan de assumpties die van tevoren zijn gesteld voor de toetsende statistiek. De uitkomsten van de respondenten waren namelijk onafhankelijk van elkaar. Ondanks dat de steekproef ($N = 34$) groot genoeg was, is er ook gekeken naar de normale verdeling van de scores op de Likert-schaal zoals getoond in Figuur 2. Hieruit is gebleken dat de verschillende scores redelijk normaal verdeeld zijn, maar dat er wel een kleine links-scheve verdeling zichtbaar is.

Figuur 2

Histogram scores Likert-schaal



Voetnoot: $N = 34$

Uit de eenzijdige matched pairs t-test bleek dat t(33)=1.783, p = 0.042 met een 95%-betrouwbaarheidsinterval van [-0.02; 0.28]. Er is naar voren gekomen dat er waarschijnlijk een verschil is bij het beoordelen van de p-waardes door studenten van de faculteit Gedrags- en Maatschappijwetenschappen, waarbij de significante wetenschappelijke artikelen beter werden beoordeeld dan niet-significante wetenschappelijke artikelen. De effectgrootte van .31 geeft aan er sprake is van een klein verschil in de beoordeling van de verschillende p-waardes.

Discussie

"In hoeverre vinden studenten van de faculteit Gedrags- en Maatschappijwetenschappen (Rijksuniversiteit Groningen) de p-waarde belangrijk voor het beoordelen van de kwaliteit van wetenschappelijke artikelen?" luidde de onderzoekvraag. Om antwoord te krijgen op deze onderzoeksvraag is een kwantitatief onderzoek uitgevoerd. Er is met een gesloten enquête via Qualtrics onderzocht hoe studenten van de faculteit Gedrags- en Maatschappijwetenschappen diverse significant en niet-significante wetenschappelijke artikelen beoordelen. Het doel van dit onderzoek was het in kaart brengen van de voorkeuren voor diverse p-waardes van studenten van de faculteit Gedrags- en Maatschappijwetenschappen om vervolgens te beoordelen of er sprake kan zijn van een voorkeur voor significante p-waarden onder deze studenten.

Binnen dit onderzoek zijn tien wetenschappelijke artikelen met vijf significante en vijf niet-significante uitkomsten voorgelegd aan de respondenten. Uit de resultaten is gebleken dat er waarschijnlijk een verschil is tussen hoe significante en niet-significante p-waardes worden beoordeeld onder studenten van de faculteit Gedrags- en Maatschappijwetenschappen. Hierbij beoordeelden de respondenten de significante p-waardes (3,89) gemiddeld hoger dan niet-significante p-waardes (3,76). De respondenten beschouwden echter geen enkele p-waarde als 'zeer onbelangrijk'. Bovendien is er sprake van een zeer klein verschil tussen de gemiddelde beoordeling van significante p-waardes en niet-significante p-waardes (slechts 0.03x hoger).

Deze resultaten sluiten aan bij de hypothese die vooraf gesteld was. Studenten, de onderzoekers van de toekomst, hebben tijdens dit onderzoek waarschijnlijk meer vertrouwen getoond in significante uitkomsten. Dit sluit aan bij de eerdere onderzoeken van Kicinski (2013) en Marin-Franch (2018) waarbij statistisch significante uitkomsten eerder werden gepubliceerd, geciteerd en/of gepromoot door onderzoekers. Bovendien komt het overeen met wat Mullins, Crowe & Wymer (2010) hebben benoemd over de invloed van publicatiebias op het gebruik van literatuur. Dit heeft namelijk een invloed op het vertrouwen in het wetenschappelijk proces onder studenten.

Bij de beoordeling door de respondenten is er wellicht sprake geweest van een Cliff-effect, echter zoals (Nelson, Rosenthal & Rosnow, 1986) veronderstellen ligt dit Cliff-effect

niet bij de kritieke waarde ($p > 0.05$), deze ligt volgens dit onderzoek bij $p < 0.02$. Er zou gezegd kunnen worden dat de respondenten een lage p-waarde als belangrijk ervaarde. Wellicht zijn studenten met de tijd kritischer geworden wat betreft de p-waarde en het gebruik van de p-waarde. Dit zou bijvoorbeeld kunnen komen door kritiek op nulhypothese toetsen (null hypothesis significance testing) (Cohen, 1994).

Beperkingen en aanbevelingen

Een discutabel punt binnen dit onderzoek gaat over de operationalisering van het woord 'belangrijk'. Respondenten kunnen een p-waarde van 0.9 'zeer belangrijk' vinden voor de beoordeling van de kwaliteit van een wetenschappelijk artikel, aangezien zij deze niet gunstig vinden voor de bevordering van de kwaliteit. Het woord belangrijk zegt in deze dus niet zo veel. Aanbevolen wordt ten eerste om het woord 'belangrijk' beter te operationaliseren, zodat er meer bewijs gevonden kan worden voor publicatiebias onder de steekproef. Wanneer er meer duidelijkheid bestaat over wat een respondent verstaat onder belangrijk, kunnen er ook correcte conclusies getrokken worden. Voor een vervolgonderzoek wordt aangeraden om na te denken over het stellen van open vragen in plaats van gesloten vragen. Op deze manier zouden de respondenten uitleg kunnen geven over hun antwoord, wat meer informatie oplevert voor het eventuele bewijs voor publicatiebias onder de steekproef.

Ten tweede kunnen er geen stellige uitspraken gedaan worden over het Cliff-effect, aangezien de p-waardes niet gelijkmatig verdeeld zijn. Gezien het kleine Cliff-effect dat zich binnen dit onderzoek laat zien rond $p > 0.02$ in plaats van rond de kritieke waarde ($p > 0.05$), zou een suggestie kunnen zijn om meer p-waarden toe te voegen rond het kantelpunt van 0.02 en 0.05 en wat minder aan de uiteinden (Nelson, Rosenthal & Rosnow, 1986). Hierdoor kan er in een vervolgonderzoek duidelijker gesteld worden op welk punt binnen de steekproef het Cliff-effect op zou kunnen treden.

Ten derde zouden er meer respondenten geworven kunnen worden, zodat de statistische power van dit onderzoek versterkt kan worden. De schattingen worden om deze reden preciezer.

In dit onderzoek behoorde ten slotte alle studenten van de faculteit Gedrags- en Maatschappijwetenschappen van de Rijksuniversiteit Groningen tot de populatie. Er werd binnen het onderzoek echter niet gevraagd tot welke richting de respondent behoorden, waardoor er niet gekeken kan worden of er een verschil aan de beoordeling van de verschillende p-waardes zat binnen de respondenten. Het zou daarom interessant kunnen zijn om te kijken naar de beoordeling van p-waardes binnen de verschillende studies van de faculteit Gedrags- en Maatschappijwetenschappen, bijvoorbeeld naar het verschil van beoordeling tussen de Academische Opleiding Leraar Basisonderwijs en Pedagogische Wetenschappen. Dit

vervolgonderzoek zou boeiend kunnen zijn, aangezien studenten van Pedagogische Wetenschappen volledig wetenschappelijk gefocust zijn en wellicht meer waarde hechten aan p-waarden ten opzichte van studenten van de Academische Opleiding Leraar Basisonderwijs, welke zowel waarde hechten aan wetenschappelijke theorie, maar daarbij altijd de ervaringen in de praktijk meenemen in hun beoordeling.

Binnen dit onderzoek is naar voren gekomen dat studenten van de faculteit Gedrags- en Maatschappijwetenschappen (Rijksuniversiteit Groningen) de p-waardes belangrijk vinden voor het beoordelen van de kwaliteit van wetenschappelijke artikelen en hierbij hebben de studenten een significante voorkeur voor significante p-waardes ten opzichte van niet-significante p-waardes. Het zou goed zijn om de wetenschappers van de toekomst het belang van zowel significante als niet-significante p-waardes op het hart te drukken, zodat de kritische blik (die geacht wordt van wetenschappers) niet tenietgedaan wordt door publicatiebias.

Referentielijst

Cohen, J. (1994). The earth is round (p < .05). *American Psychologist, 49*(12), 997–1003. https://doi.org/10.1037/0003-066X.49.12.997

Demaria, A. N. (2004). Publication bias and journals as policemen. *Journal of the American College of Cardiology*, *44*(8), 1707–8.

Fanelli, D. (2010). "Positive" results increase down the hierarchy of the sciences. PloS one, 5(4), e10068.

Ferguson, C. J., & Brannick, M. T. (2012). Publication bias in psychological science: Prevalence, methods for identifying and controlling, and implications for the use of meta-analyses. Psychological Methods, 17, 120-128. doi: 10.1037/a0024445

Franco, A., Malhotra, N., & Simonovits, G. (2014). Publication bias in the social sciences: unlocking the file drawer. *Science*, *345*(6203), 1502–1505.

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

Ioannidis, J.P.A. (2005). Why Most Published Research Findings Are False. PLoS Med 2(8): e124. https://doi.org/10.1371/journal.pmed.0020124

Kicinski, M. (2013). Publication bias in recent meta-analyses. *Plos One*, *8*(11), 81823. https://doi.org/10.1371/journal.pone.0081823

Marin-Franch, I. (2018). Publication bias and the chase for statistical significance. *Journal of Optometry*, *11*(2), 67–68.

Miller, S. C., & Moulder, J. E. (1998). Publication of negative results is an essential part of the scientific process. *Radiation Research*, *150*(1), 1–2.

Moss, J., & De Bin, R. (2021). Modelling publication bias and p-hacking. *Biometrics*, *(20210912)*. https://doi.org/10.1111/biom.13560

Mullins, M., Crowe, E. A., & Wymer, C. (2015). A question of trust: publication bias and student views of psychological literature. *North American Journal of Psychology*, *17*(1), 59–76.

Nelson, N., Rosenthal, R., & Rosnow, R. L. (1986). Interpretation of significance levels and effect sizes by psychological researchers. *American Psychologist*, *41*(11), 1299–1301. https://doi.org/10.1037//0003-066X.41.11.1299

Rockwell, S., Kimler, B. F., & Moulder, J. E. (2006). Publishing negative results: the problem of publication bias. *Radiation Research*, *165*(6), 623–625. https://doi.org/10.1667/RR3573.1

Rosenthal, R., & Gaito, J. (1963). The interpretation of levels of significance by psychological researchers. *The Journal of Psychology*, *55*(1), 33–38. https://doi.org/10.1080/00223980.1963.9916596

Rosenthal, R. (1979). The file drawer problem and tolerance for null results. *Psychological Bulletin*, *86*(3), 638–641. https://doi.org/10.1037/0033-2909.86.3.638

Torgerson, C.J. (2006) PUBLICATION BIAS: THE ACHILLES' HEEL OF SYSTEMATIC REVIEWS?, British Journal of Educational Studies, 54:1, 89-102, DOI: 10.1111/j.1467-8527.2006.00332.x

Van Aert, R. C. M., Wicherts, J. M., van Assen, M. A. L. M., & Macleod, M. R. (2019). Publication bias examined in meta-analyses from psychology and medicine: a meta-meta-analysis. *Plos One*, *14*(4). https://doi.org/10.1371/journal.pone.0215052

Bijlagen

Voorbeeld artikelen enquête

Deze bijlage bevat een voorbeeld van één artikel inclusief de vragen die hierbij gesteld zijn afkomstig uit de enquête. Daarna volgen de overige negen artikelen uit het eerste block. Voor de volledige enquête kan de volgende link gebruikt worden: https://rug.eu.qualtrics.com/jfe/preview/SV_41JtV7uZLDmOfUG?Q_CHL=preview&Q_SurveyVersionID=current

1/10) Het effect van de COVID-19 lockdown op de mentale gezondheid van jongeren.
Uitgebracht in 2021

Inleiding: Onderzoekers willen weten wat de mentale impact van de COVID-19 crisis is op jongeren, een groep waarvan het mentaal lijden door velen onderschat werd.

Methode: Er is gebruik gemaakt van de data van de Center for Epidemiological Studies Depression Scale (CES-D-8) om de associatie te testen tussen depressieve symptomen en lockdowngerelateerde factoren die werden weerhouden in een recente overzichtsstudie uit tijdschrift The Lancet. Daarnaast werd er bij een andere groep de General Health Questionnaire (GHQ12) afgenomen. Deze testen zijn afgenomen bij in totaal 20.000 jongeren.

Resultaten: Ongeveer twee derde (65,49%) van de deelnemers vertoonde significante mentale stress, gedefinieerd als een score van hoger dan 3 op de GHQ-12 ($\alpha = 0.69$, $p = 0.03$) en gelijk aan een verscheidenheid aan verontrustende psychologische symptomen. De impact van de lockdown op jongeren lijkt vrij indrukwekkend in beide studies. Dit kan mogelijk verklaard worden door een verhoogde nood aan interactie met leeftijdsgenoten en sociale stimuli in deze leeftijdsgroep, hetgeen aan banden gelegd wordt tijdens een lockdown.

Vraag 1: In hoeverre vindt u de steekproefgrootte (20.000) van belang voor het beoordelen van de kwaliteit van dit artikel?

- 1 – zeer onbelangrijk
- 2 – onbelangrijk
- 3 – neutraal
- 4 – belangrijk
- 5 – zeer belangrijk

Vraag 2: In hoeverre vindt u de cronbach's alfa (0.69) van belang voor het beoordelen van de kwaliteit van dit artikel?

- 1 – zeer onbelangrijk
- 2 – onbelangrijk
- 3 – neutraal
- 4 – belangrijk
- 5 – zeer belangrijk

Vraag 3: In hoeverre vindt u de p-waarde (0.03) van belang voor het beoordelen van de kwaliteit van dit artikel?

- 1 – zeer onbelangrijk
- 2 – onbelangrijk
- 3 – neutraal
- 4 – belangrijk
- 5 – zeer belangrijk

Vraag 4: In hoeverre vindt u het uitgavejaar (2021) van belang voor het beoordelen van de kwaliteit van dit artikel?

- 1 – zeer onbelangrijk
- 2 – onbelangrijk
- 3 – neutraal
- 4 – belangrijk
- 5 – zeer belangrijk

Vraag 5: In hoeverre zou u dit artikel voor eigen gebruik gebruiken?

- 1 – helemaal niet gebruiken
- 2 – niet gebruiken
- 3 – neutraal
- 4 – gebruiken
- 5 – absoluut gebruiken

2/10) Bij een jeugdpaniekstoornis in een gerandomiseerde klinisch onderzoek is er geen verschil tussen groeps- of individuele behandeling.

Uitgebracht in 2008

Inleiding: De Randomised Clinical Trial (RCT) en Cognitieve Gedragstherapie (CGT) hebben de status van een empirisch ondersteunde behandeling terwijl het niet altijd effectief blijkt. We weten echter weinig over de effectiviteit van alternatieve behandelingen. Voor dit onderzoek wordt een individuele versus een groepsformat van de manuele cognitieve gedragstherapie (FRIENDS) vergeleken qua effectiviteit bij kinderen met een angststoornis.

Methode: 127 kinderen gediagnosticeerd met een verlatingsangststoornis (gemeten met een likertschaal, $\alpha = 0.81$) zijn willekeurig verdeeld over een individuele en een groepsbehandeling. Analyses werden getrokken op basis van een chi-kwadraadtoets en regressieanalyses.

Resultaten: 48% van de kinderen met individuele behandeling versus 41% van de kinderen met groepsbehandleing hadden geen last meer van een angststoornis na hun behandeling. Regressieanalyses hebben geen significante verschillen aangetoond tussen de individuele en groepsbehandeling. Kinderen lieten bij beide behandelingen verbetering zien door hun behandeling, waarbij er geen verschil zat tussen groeps- of individuele behandeling ($p = 0.23$).

3/10) De invloed van het geslacht van de docent op de effectiviteit van verschillende leesmotivatie-interventies.
Uitgebracht in 2019

Inleiding: Kinderen uit Nederland scoren hoog op leesvaardigheid maar laag op leesmotivatie. Daarom is het van belang om te onderzoeken wat de effectiviteit van verschillende leesmotivatie-interventies is.

Methode: De interventie werd getest in de huidige onderwijssetting, met een voor- en nameting. Hierbij werd met behulp van een vragenlijst ($\alpha = 0.76$) aan 250 leerlingen gekeken of de motivatie hoger is geworden.

Resultaten: Er bleek een verschil te zijn in leesmotivatie, voor en na de interventie, maar deze

bleek niet significant (p=0.54). Ook werd de interventie ondersteund door middel van leesvaardigheden te verbeteren, dit zou van invloed kunnen zijn op de uitkomsten.

4/10) De invloed van geslacht op de aard van conflicten in een klas.
Uitgebracht in 2009

Inleiding: Een onderzoek naar het verschil in de aard van conflicten tussen mannelijke en vrouwelijke docenten met hun leerlingen.

Methode: Data wordt door middel van een vragenlijst vergaard. Er doen 100 docenten mee die allen in groep 6 lesgeven. De aard van een conflict wordt gemeten aan de hand van een vragenlijst waar hogere scores conflicten zijn met meer agressief taalgebruik ($\alpha = 0.58$). De analyse wordt met behulp van een regressieanalyse gedaan.

Resultaten: Er is een verschil gevonden in de aard van conflicten in de klas tussen mannelijke en vrouwelijke docenten ($p = 0.002$). Bij vrouwelijke docenten zijn er conflicten gerapporteerd met meer agressief taalgebruik dan bij mannelijke docenten.

5/10) In hoeverre heeft de leeftijd van een docent invloed op de schoolprestaties van allochtone leerlingen?
Uitgebracht in 2014

Inleiding: Een onderzoek waarbij wordt gekeken of de leeftijd van een docent van invloed is op de schoolprestaties van 6000 specifieke allochtone leerlingen. Allochtone leerlingen hebben vaak een onderwijsachterstand en het is van belang te onderzoeken of de leeftijd van de docent hier invloed op kan hebben.

Methode: De onafhankelijke variabele is schoolprestaties van allochtone leerlingen, dit is geoperationaliseerd als de scores op een gestandaardiseerde toets ($\alpha = 0.67$).

Resultaten: Er is een verband gevonden tussen de schoolprestaties van allochtone leerlingen en de leeftijd van de docent ($p=0.02$). Naarmate de leeftijd van de docent toeneemt dalen de schoolprestaties van allochtone leerlingen.

6/10) De Invloed van Motivatie, Behoefte Ondersteuning en Prestatie-emoties op de Academische Prestaties van hbo-studenten.
Uitgebracht in 2016

Inleiding: Hogescholen hebben te maken met hoge uitvals- en lage slagingspercentages, wat schadelijk is voor de instellingen maar ook voor studenten.

Methode: Door middel van een enquête met gesloten vragen is er onderzoek gedaan naar verschillende factoren die invloed kunnen hebben op de academische prestaties van hbo-studenten ($\alpha = 0.83$). Tijdens het onderzoek zijn 1500 hbo-studenten betrokken.

Resultaten: Uit het onderzoek bij 1500 hbo-studenten is gebleken dat prestatie-emoties van significant voorspellende waarde zijn van het studieresultaat ($p=0.38$). Motivatie en de mate van behoefte ondersteuning blijken in de steekproef, tegen de verwachting in, geen significante voorspellende waardes.

7/10) De mate van invloed van groepscohesie en sociale veiligheid op participatie in de klas.
Uitgebracht in 2019

Inleiding: Dit artikel heeft als doel te onderzoeken of de interpretatie van leerlingen over sociale veiligheid en groepscohesie voorspellers zijn voor de participatie van de leerling. Een slechte sociale veiligheid of lage mate van groepscohesie kan desastreus zijn voor de participatie van leerlingen in de klas.

Methode: Op een meetmoment zijn er meerdere vragenlijsten aangeboden (alle α's boven 0.62). Er deden 64 leerlingen mee aan de vragenlijsten. De resultaten zijn getest op normaliteit door middel van een histogram, boxplot en een Kolmogorov-Smirnov analyse. De data is geanalyseerd door een Spearman's rho correlatie. De data bleek niet allemaal normaal verdeeld waardoor er non-parametrisch getest moest worden.

Resultaten: Er zijn geen significante verbanden gevonden tussen sociale veiligheid en participatie, en groepscohesie en participatie (alle p-waarden > 0.069). De nulhypothese kan niet verworpen worden aan de hand van de resultaten.

8/10) Stress bij basisschoolleerlingen.
Uitgebracht in 1990

Inleiding: Een onderzoek naar stress en coping in schoolsituaties die door leerlingen uit groep zeven en acht als belastend worden waargenomen. Het is een onderwijspsychologische studie die tot doel heeft een bijdrage te leveren aan het inzicht in stress en coping als sociaal-affectieve aspecten van het onderwijsleerproces.

Methode: 206 jongens en 220 meisjes uit groep zeven en acht namen deel aan het onderzoek. 336 leerlingen werden een open vragenlijst ($\alpha = 0.58$) aangeboden en 90 leerlingen hebben een individueel interview afgelegd.

Resultaten: De algemene conclusie is, dat jonge kinderen die recent een dramatische levensgebeurtenis hebben meegemaakt ($p = 0.166$), meer aanpassingsproblemen op school ervaren dan leeftijdsgenootjes die een dergelijke crisis niet hebben doorgemaakt.

9/10) Invloed van Mindfulness op burn-out van docenten op het middelbaar onderwijs.
Uitgebracht in 2011

Inleiding: Er wordt in dit artikel gekeken naar de invloed van Mindfulness op burn-out van docenten in het middelbaar onderwijs.

Methode: Bij 220 leraren zijn de 'Five Facets Mindfulness Questionnaire (FFMQ) en de Burn-out Assessment Tool (BAT) afgenomen (alle $\alpha > 0.82$). De vragen zijn gemeten op een Likert-schaal en vervolgens getoetst met correlatie- en regressieanalyses.

Resultaten: Er blijkt een negatieve relatie te zijn tussen mindfulness en burn-out ($ß = -.19$, $p = .049$). Hoe hoger de docenten scoorden op de FFMQ, hoe lager ze scoorden op de BAT.

10/10) Invloed van problematisch cannabisgebruik op het psychisch welzijn van studenten.
Uitgebracht in 2018

Inleiding: Er wordt in dit onderzoek gekeken naar de invloed van problematisch cannabisgebruik op het psychisch welzijn van studenten.

Methode: Door 7181 studenten is de General Health Questionnaire (GHQ) ($\alpha = 0.87$) ingevuld. Daarnaast werd een korte algemene vragenlijst ingevuld over het cannabisgebruik. Vervolgens is de data geanalyseerd met SPSS.

Resultaten: Er bleek een significant verband tussen problematisch cannabisgebruik en het psychisch welzijn van studenten. Het bleek dat studenten die meer gebruik maakten van cannabis, lager scoorden op de GHQ en dus een minder psychisch welzijn hadden ($p=0.001$).