

Empirische basis van conclusies

Handvatten voor de empirische taalonderzoeker

Anita Eerland & Huub van den Bergh

TT 38 (2): 139–146

DOI: 10.5117/TVT2016.2.EERL

Abstract

Empirically sound conclusions: A guide for language researchers

Recently, there has been (and still is) a lot of discussion about the replicability of scientific findings. Hornikx and Batenburg (2016) try to explain the non-replicability in science, and discuss the situation for language research. We make the point that empirically sound conclusions will contribute to the replicability of science. In this commentary, we discuss several practices that will help the empirical language researcher to draw more solid conclusions from their data.

Keywords: replicability, experimental design

1 Inleiding

Het onderwerp dat Hornikx en Batenburg (2016) aansnijden, repliceerbaarheid van onderzoeksresultaten, is uitermate belangrijk. De repliceerbaarheid van onderzoek is essentieel in onze zoektocht naar kennis. Dat onderzoeksresultaten regelmatig niet-repliceerbaar zijn gebleken is problematisch. Deze niet-repliceerbaarheid kan natuurlijk veroorzaakt worden door frauduleuze onderzoekers, zoals Stapel (zie Hornikx & Batenburg, 2016). Echter, men dient uitermate terughoudend te zijn een non-replicatie te zien als bewijs van fraudeleus handelen. Er zijn namelijk tal van alternatieve verklaringen mogelijk voor de niet-repliceerbaarheid van onderzoeksresultaten. Het trekken van generaliseerbare conclusies uit de onderzoeksresultaten is immers niet eenvoudig. Vaak bestuderen we relaties tussen variabelen, maar trekken we conclusies over relaties tussen concep-

ten. Dit leidt tot een *mismatch* tussen de claims die worden gemaakt in wetenschappelijke artikelen en de empirische basis hiervan (Clark, 1973; Meuffels & Van den Bergh, 2005, 2006). Een empirisch onderzoek wordt namelijk altijd uitgevoerd binnen een bepaalde context (bijvoorbeeld: de steekproef komt uit een specifieke populatie (en is lang niet altijd aselekt), er worden een specifieke taak en specifieke materialen gebruikt, het onderzoek vindt plaats in een specifieke tijd en omgeving). We mogen er daarom niet zonder meer op vertrouwen dat de bevindingen generaliseerbaar zijn en dat elke specifieke operationalisatie van een construct zonder meer opgevat kan worden als zijnde een valide operationalisatie. Een voor de hand liggende oplossing voor de bestaande mismatch lijkt het trekken van meer specifieke conclusies. In dit artikel beargumenteren we echter dat het vergroten van de repliceerbaarheid van onderzoek een betere aanpak is. Hiervoor zullen we de empirische taalonderzoeker enkele handvatten aanreiken.

2 Directe en conceptuele replicaties

Conclusies over de relatie tussen concepten zijn vaak gebaseerd op slechts een enkele bevinding binnen een beperkte onderzoekssetting. Hierdoor wordt de suggestie gewekt dat we een oplossing voor het probleem met de generaliseerbaarheid van onderzoeksresultaten dienen te zoeken in het uitvoeren van replicatiestudies. Recent verscheen in *Science* het resultaat van 100 replicatiestudies (Open Science Collaboration, 2015). De conclusie is dat de resultaten van de meeste originele studies niet repliceerbaar blijken. Ook bleken de effectgroottes van de replicaties half zo groot als die van de originele studies. Deze conclusies liegen er niet om. Toch dienen we ook hier weer voorzichtig te zijn in de interpretatie van deze bevindingen. We moeten namelijk rekening houden met de natuurlijke variatie in effectgroottes. Elke originele studie werd slechts eenmaal gerepliceerd. Wanneer de originele studie wel en de replicatie geen effect vindt, kunnen we simpelweg niet concluderen dat een effect niet (of wel) bestaat.

Een goed initiatief dat inspringt op de tekortkomingen van een enkele replicatie zijn de grootschalige replicatieprojecten, zoals de *registered replication reports*. De veronderstelling dat het repliceren van een bepaalde studie volgens een vooraf afgesproken protocol een generaliserend antwoord oplevert door deze studies mee te nemen in een meta-analyse (Hornikx & Batenburg, 2016) dient echter enigszins genuanceerd te worden. Het gaat hier namelijk om replicaties die zoveel mogelijk lijken op de

originele studie. Dat wil zeggen dat de replicatiestudies gebruik maken van dezelfde materialen, taken en procedure als de originele studie en dat de steekproef uit een gelijkende populatie wordt getrokken. We noemen dergelijke studies dan ook directe replicaties¹ (Schmidt, 2009). Een meta-analyse waarin directe replicaties worden meegenomen, zorgt voor een grotere mate van vertrouwen in de bevindingen. Echter, hoe directer een replicatie, hoe minder generaliseerbaar de resultaten zijn.

Hornikx en Batenburg (2016) noemen de studie van Eerland et al. (2016) als mooi voorbeeld voor de taalbeheersing van een grootschalig replicatieproject. Verschillende onderzoeksgroepen hebben in totaal twaalf directe replicaties uitgevoerd van een studie naar de invloed van *grammatical aspect* op de mate waarin een criminele actie als intentioneel wordt beschouwd (verder omschreven als *intentionality*; Hart & Albarracín, 2011). De meta-analyse, waarin alleen de replicaties werden meegenomen die in een laboratorium waren uitgevoerd, liet geen effect zien van *grammatical aspect* op *intentionality*. Toch kunnen we op basis van deze bevindingen niet concluderen dat *grammatical aspect* geen invloed heeft op hoe mensen denken over de intentie van een actie. Wel kunnen we met redelijke zekerheid stellen dat *grammatical aspect* in deze specifieke studie, met dit specifieke design en vignet en deze specifieke steekproef geen invloed heeft op de mate waarin mensen een actie als intentioneel beschouwen. De directe replicaties bieden onvoldoende basis voor het maken van meer algemene claims.

Om middels replicaties de mismatch tussen bevindingen en claims te verkleinen, dienen vooral zogenaamde conceptuele replicaties te worden uitgevoerd. Een originele studie en een conceptuele replicatie testen eenzelfde hypothese of experimenteel resultaat (Schmidt, 2009). Ze zijn gebaseerd op dezelfde theorie, maar onderzoeken de relatie tussen concepten op een andere manier. Zo kan een conceptuele replicatie afwijken van de originele studie qua design, steekproef en meetinstrumenten (zie Louwerse, Hutchinson, Tillman & Recchia (2015) voor een mooi voorbeeld van een meta-analyse van conceptuele replicaties).

Zowel directe als conceptuele replicaties zijn informatief, maar elk op hun eigen manier. Het is belangrijk om de voor- en nadelen van beide soorten replicaties te kennen, zodat je uit hun bevindingen de juiste conclusies kunt trekken.

3 Replicaties binnen een studie

Replicaties kunnen ook binnen een studie uitgevoerd worden. We spreken in dergelijke gevallen dan vaak niet van replicaties maar van kruisvalidatie. Gedacht kan worden aan kruisvalidatie op cruciale variabelen, kruisvalidatie op de steekproef, of kruisvalidatie van het effect van de manipulatie.

Kruisvalidatie op variabelen ligt dicht tegen een *multiple message design* aan. Met kruisvalidatie op de variabelen wordt aangenomen dat met elk instrument altijd iets gemeten wordt wat kenmerkend is voor dat specifieke meetinstrument, iets dat kenmerkend is voor deze specifieke operationalisatie van het construct in kwestie. Omdat elk instrument deels iets specifiek meet, is het verstandig cruciale variabelen op verschillende manieren te operationaliseren, zodat nagegaan kan worden of de verwachte relaties voor de verschillende operationalisaties aangetoond kunnen worden. Mooie voorbeelden hiervan komen uit het intelligentie-onderzoek. Niet voor niets zei Elshout (1976) dat elke intellectuele vaardigheid op ten minste drie verschillende manieren geoperationaliseerd dient te worden. Helaas nemen we dit zelden ter harte, hoewel het een relatief eenvoudige ingreep is. Bijvoorbeeld: Van Dooren, Evers-Vermeul en Van den Bergh (2015) demonstreerden bij verschillende teksten dat effecten van tekststructuur op begrip van lezers voornamelijk aan te tonen zijn bij slechtere lezers. Omdat zij vier verschillende teksten gebruikten kon het effect van tekststructuur op tekstbegrip in deze studie viermaal aangetoond worden. Het gevolg van deze kruisvalidatie op de onafhankelijke variabele is dat de getrokken conclusies minder afhankelijk geworden zijn van één specifieke operationalisatie van tekststructuur. Daarnaast hebben Van Dooren et al. het tekstbegrip van leerlingen op twee manieren geoperationaliseerd: met meerkeuzevragen en met een zogenaamde cloze-tekst. Met als resultaat dat Van Dooren et al. met enige zekerheid aannemelijk gemaakt hebben dat het positieve effect van een duidelijke tekststructuur op het begrip van minder goede lezers. In deze conclusie is zij niet meer gebonden aan één specifieke tekst of één specifieke operationalisatie van tekstbegrip.

Kruisvalidatie op de steekproef. In vrijwel alle empirische studies wordt gebruikgemaakt van een steekproef. Nagegaan wordt of een verwacht effect in deze steekproef aangetoond kan worden. Zelden echter beschikken we over een echt aselechte steekproef. Het is wenselijk om na te gaan of gehypothetiseerde effecten aangetoond kunnen worden in verschillende substeekproeven. Kan bijvoorbeeld het gehypothetiseerde effect aangetoond worden wanneer de steekproef aselekt opgedeeld wordt in twee substeekproeven? Of, kan het effect ook aangetoond worden bij verschil-

lende steekproeven als mannen en vrouwen of jongeren en ouderen. He-
laas zijn dergelijke analyses schaars, hoewel ze de reikwijdte van conclu-
sies kunnen vergroten. Kuhlemeier (1996) liet bijvoorbeeld zien dat lees-
vaardigheid gemeten bij vmbo'ers niet exact hetzelfde is als leesvaardig-
heid gemeten bij vwo'ers, ook al zijn bij beide groepen leerlingen exact
dezelfde toetsen afgenomen. Het gevolg is dat de relatie met andere varia-
belen verschilt voor vmbo- en vwo-leerlingen. Met een dergelijke kruis-
validatie op de steekproef kunnen dus meer genuanceerde conclusies ge-
trokken worden; conclusies die (hopelijk) robuuster zijn dan wanneer al-
leen naar de steekproef als geheel gekeken wordt.

Kruisvalidatie op de manipulatie. In experimenteel onderzoek wordt het
effect van een manipulatie getoetst. Echter, in veel gevallen wordt dit effect
éénmaal getoetst, zoals in het klassieke pretest-posttest-controle-groep-ont-
werp. Recentelijk komen ook meer geavanceerde onderzoeksontwerpen in
zwang, zoals het *crossed lagged panel design*. Dit onderzoeksontwerp is
gebruikt in een recent onderzoek naar evaluatie van een nieuwe lesme-
thode voor schrijfvaardigheid: Tekster (Koster & Bouwer, 2016). In deze
studie is de schrijfvaardigheid van leerlingen op drie momenten gemeten
is. In de ene groep vond de interventie plaats tussen de eerste en de tweede
meting. In de andere groep vond de interventie tussen de tweede en de
derde meting plaats. In één onderzoek kan derhalve het effect van de
interventie tweemaal getoetst worden. Bovendien kan nagegaan worden
in hoeverre het effect van de interventie afhankelijk is van kenmerken van
beide groepen, waarmee een deel van de interne validiteit van een onder-
zoek onderwerp van statistische toetsing wordt. In dit *crossed lagged panel
design* wordt het effect van een manipulatie tweemaal getoetst, en wordt
minder gekapitaliseerd op steekproeftoevalligheden.

Een geheel andere manier om tot robuustere conclusies te komen wordt
ook in de bijdrage van Hornikx en Batenburg (2016) genoemd: zogenaamde
multiple message designs. Feitelijk wordt hiermee onderkend dat we in
onderzoek vaak gelijktijdig willen generaliseren naar zowel de populatie
van personen als de populatie van stimuli. Dat betekent dat we in één
analyse zowel rekening moeten houden met de variatie tussen personen
als de variatie tussen stimuli. Mooie voorbeelden hiervan kunnen ontleend
worden aan onderzoek naar effecten van vraagformulering. Werd vroeger
het antwoord op één vraag in een positieve variant vergeleken met het
antwoord op dezelfde vraag in een negatieve variant (zie bijvoorbeeld:
Rugg, 1941), in een *multiple message design* wordt de variantie van het
effect (van vraagformulering) gekwantificeerd. Uit diverse onderzoeken
blijkt een duidelijk verschil tussen positief en negatief geformuleerde vra-

gen, waarbij het antwoord op de negatieve vraag positiever is. Echter, de grootte van dit effect verschilt van vraag tot vraag, bij sommige vragen is het effect veel groter en bij andere veel kleiner, en bij enkele vragen blijkt het verschil in antwoorden zelfs omgedraaid (zie bijvoorbeeld: Holleman, 1996; Kamoen, 2012). De winst met een *multiple message design* is dat we niet alleen een schatting krijgen van het gemiddelde effect, maar ook een indicatie van de verdeling van het effect over verschillende stimuli (of in dit voorbeeld: enquêtevragen).

Meta-analyses zijn geweldig handig om een overzicht te krijgen van de stand van zaken in een bepaald vakgebied. Hoewel meta-analyses vaak opgevat worden als mogelijke vervanging van een *multiple message designs* moet opgemerkt worden dat dit onterecht is. In een meta-analyse worden vaak effectgroottes (ES) van het verschil tussen condities als afhankelijke variabele beschouwd ($ES = [\bar{X} - \bar{Y}] / SD_{\text{pooled}}$). Met een meta-analyse van een aantal *single message studies* kan dan ook geen rekening gehouden met de verschillen tussen stimuli (en die kunnen relatief groot zijn). Het gevolg is dat de vraag of een gevonden effect gelijktijdig gegeneraliseerd kan worden van de steekproef personen naar de populatie waaruit deze afkomstig zijn, als van de steekproef van stimuli naar de populatie stimuli moet helaas onbeantwoord blijven. Een en ander betekent niet dat aan het nut van meta-analyses getwijfeld moet worden, maar wel dat zij niet kunnen dienen als vervanging voor de broodnodige *multiple message designs*.

Kortom, repliceerbaarheid van empirisch onderzoek is een uitermate belangrijk issue. Met dit onderwerp in het achterhoofd zijn er gelukkig nog tal van manieren waarop wij ons onderzoek kunnen verbeteren. Enkelen hiervan hebben wij hierboven aangestipt, maar er zijn natuurlijk nog veel meer manieren om de generaliseerbaarheid te vergroten, zowel met behulp van sterkere onderzoeksontwerpen, als met meer geavanceerde statistische methoden.

Noot

1. Soms spreekt men ook van exacte replicaties. Echter exacte replicaties zijn niet mogelijk. Men kan immers niet gebruik maken van precies dezelfde proefpersonen en ook het exacte moment en de omstandigheden van de originele testafname kunnen niet worden teruggehaald.

Referenties

- Clark, H.H. (1973). The language-as-fixed-effect fallacy. *Journal of Verbal Learning and Verbal Behavior*, 12, 335-359.
- Eerland, A., Sherrill, A.M., Magliano, J.P., Zwaan, R.A., Arnal, J.D., Aucoin, P., ... Prenoveau, J.M. (2016). Registered replication report: Hart & Albarracín (2011). *Perspectives on Psychological Science*, 11 (1), 158-171.
- Elshout, J.J. (1976). *Karakteristieke moeilijkheden in het denken*. Amsterdam: Universiteit van Amsterdam.
- Hart, W., & Albarracín, D. (2011). Learning about what others were doing: Verb aspect and attributions of mundane and criminal intent for past actions. *Psychological Science*, 22 (2), 261-266.
- Holleman, B., (2000). *The forbid/allow asymmetry: on the cognitive mechanisms underlying wording effects in surveys*. Amsterdam: Rodopi.
- Hornikx, J., & Batenburg, A. (2016). Integriteit in kwantitatief, empirisch onderzoek: problemen en mogelijke oplossingen. *Tijdschrift voor Taalbeheersing*, 38 (2), 119-131.
- Kamoen, N. (2012). *Positive versus negative: A cognitive perspective on wording effects for contrastive questions in attitude surveys*. Utrecht: LOT.
- Kamoen, N., Holleman, B., & Bergh, H. van den (2007). Hoe gemakkelijk is een niet moeilijke tekst: Een meta-analyse naar het effect van vraagformulering in tekstevaluatieonderzoek. *Tijdschrift voor Taalbeheersing*, 29 (4), 314-332.
- Koster, M., & Bouwer, R. (2016). *Bringing writing research into the classroom: The effectiveness of Tekster, a newly developed writing program for elementary students*. Utrecht: Universiteit Utrecht.
- Kuhlemeier (1996). *Taalvaardigheid, taalactiviteiten en taalattitudes: Een validatiestudie*. Cito: Arnhem.
- Louwerse, M.M., Hutchinson, S., Tillman, R., & Recchia, G. (2015). Effect size matters: The role of language statistics and perceptual simulation in conceptual processing. *Language, Cognition and Neuroscience*, 30 (4), 430-447.
- Meuffels, B., & Bergh, H. van den (2005). De ene tekst is de andere niet: The language-as-a-fixed-effect fallacy revisited: Methodologische implicaties. *Tijdschrift voor Taalbeheersing*, 27 (2), 106-125.
- Meuffels, B., & Bergh, H. van den (2006). De ene tekst is de andere niet: The language-as-a-fixed-effect fallacy revisited: Statistische implicaties. *Tijdschrift voor Taalbeheersing*, 28 (4), 323-345.
- Open Science Collaboration (2015). Estimating the reproducibility of psychological science. *Science*, 349 (6251), aac4716, DOI: 10.1126/science.aac4716.
- Rugg, D. (1941). Experiments in wording questions II. *Public Opinion Quarterly*, 5, 91-92.
- 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 (2), 90-100.

Over de auteurs

Anita Eerland is post doctoraal onderzoeker aan de Universiteit Utrecht bij het Utrechts Instituut voor Linguïstiek OTS.

E-mail: a.eerland@uu.nl

Huub van den Bergh is hoogleraar toetsing en didactiek van taalvaardigheid aan de Universiteit Utrecht bij het Departement TLC en het Utrechts Instituut voor Linguïstiek (UIL).

Email: H.vandenBergh@uul.nl