Vertrouwen is goed, kwetsbaarheid is beter: over gevoeligheidsanalyse in onderzoek

De online Dikke van Dale kent het woord ‘vertrouwenswaardig’ niet. Evengoed gebruikt de Onderzoeksraad voor veiligheid het woord in dit filmpje over belangenconflicten in wetenschappelijk onderzoek. In een geweldige TEDtalk over maatschappelijk vertrouwen legt Onora O’Neill uit dat we niet domweg ons vertrouwen moeten herwinnen, maar in onze eigen trustworthiness moeten investeren. En *trustworthy* worden we door ons kwetsbaar op te stellen. In haar sokkenwinkel mag je daar gekochte sokken terugbrengen als die niet bevallen. Hup, geld terug, geen vragen. Bij zo’n winkel koop je graag.

Maar hoe werkt vertrouwen in wetenschappelijk onderzoek? In mijn werk voor HvA-FGSB’s Open Science Support Desk probeer ik onderzoekers te bewegen kwetsbaarheid te omarmen door o.a. zoveel mogelijk onderzoeksmaterialen te delen. Denk aan vragenlijsten, analysecode, de data etc. Behalve door transparantie kun je je verder kwetsbaar opstellen door gevoeligheidsanalyse. Het doel van gevoeligheidsanalyse is nagaan wat bronnen zijn van variatie in resultaten wanneer je aannames en keuzes in je onderzoek varieert. Een voorbeeld. Je laat een waarneming (of, bijvoorbeeld in een meta-analyse, een gehele studie) weg en kijkt wat die weglating doet met je *overall* resultaat. Op die manier zie je of het resultaat (teveel) afhangt van één waarneming (of studie). In veel statistische software zijn zulke techniekjes voorgekookt en makkelijk uit te voeren. Ze niet gebruiken is verwijtbaar. Ander voorbeeld: in follow-up-onderzoek naar het voorkómen van nare (gezondheids)-toestanden kun je voor alle personen waarvan je niet kon achterhalen of ze in die nare toestand terecht kwamen, doen alsof dat wel het geval is en dan opnieuw gaan rekenen. In jargon heet dat ‘assuming the worst’. Als je resultaat dat ‘overleeft’, is het (op dat punt) robuust.

Ik haat artikelen waarin ik na uren turen in teksten en tabellen, in het discussiedeel word aangemoedigd de resultaten met voorzichtigheid te interpreteren. Wat een laffe manier van schrijven. Vertel me liever waarom die voorzichtigheid gewenst is. Veel onderzoekers doen dat overigens wel, maar vervelen me met beperkingen (limitations) die ik al op 100 kilometer zag aankomen: “The small sample size of our study limited the precision of our estimates.” Val me niet lastig met dat soort schijn(h)eerlijkheid! Vertel me over de beperkingen van de vertrouwenswaardigheid van je resultaten die ik, als lezer, níet gemakkelijk zelf kan verzinnen.

Wat is eigenlijk een beperking? Ik zie tenminste vier types: type-1 is opgelegd door de vraagstelling (opgelegd), type-2 is een gevolg van de gekozen methode (onvermijdelijk), type-3 vloeit voort uit onvolkomenheden in de uitvoering (vermijdbaar), en type-4 hangt samen met keuzes in je data-analyse (direct onderzoekbaar). Een

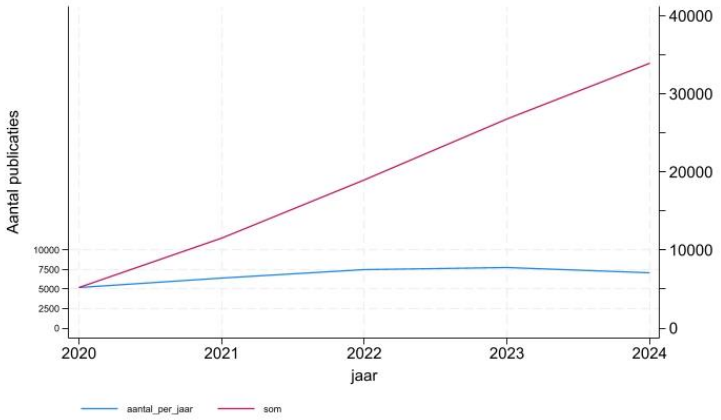
Figuur 1- typen beperkingen

voorbeeld van een (opgelegde) type-1 beperking is riskante leefwijzen. Stel je wilt weten of langdurig kettingroken bloedkanker veroorzaakt bij mensen. Experimenteel onderzoek met at random toewijzen van verplicht 40 sigaretten roken per dag gedurende tenminste 20 jaar is verboden. Het zal zonder randomisatie moeten en daaruit vloeien beperkingen voort voor de vertrouwenswaardigheid van de onderzoeksresultaten. Type-2 beperkingen gaan om *onvermijdelijke* beperkingen die voortvloeien uit bewuste keuzes, bijvoorbeeld de keuze voor live focusgroepen in plaats van online persoonlijke interviews. Je weet niet of een verlegen persoon in de focusgroep iets belangrijks en unieks had gezegd als je haar persoonlijk had geïnterviewd. Type-3 gaat over *onbedoelde* en niet verwaarloosbare gevolgen van zaken die niet volgens plan verliepen, protocolschendingen. In mijn promotie-onderzoek verbrak een fysiotherapeut de blindering van de therapie door bewust de ultrageluid-transducer in een bakje water te houden. Type-4, tenslotte, betreft keuzes die je maakt in de data-analyse. In een groot onderzoek naar het effect van preventief huisbezoek op het valrisico bij ouderen, sluit je de drie enige vrouwen met een migratie-achtergrond uit van de analyse om nadrukkelijk niet de indruk te wekken dat je resultaten ook gelden voor personen uit deze subgroep. Ter zake deskundige lezers kunnen de ernst van beperkingen van het type-1 en 2 meestal zelf beoordelen. Maar zonder kwetsbare transparantie (en toegang tot de data) is het opsporen en wegen van type 3 en 4 beperkingen bijna onmogelijk.

Ik vind het belangrijk om lezers inzicht te geven in de gevolgen van type-4 (en deels van type 3) beperkingen van onderzoek. Nu vond er de laatste 10-15 jaar een stille revolutie plaats in de methodes voor gevoeligheidsanalyse. We kennen nu multiverse-, vibration of effects-, en specificatiecurve analyse. Het gemeenschappelijke idee is dat je álle geldige, niet-redundante analyses uitvoert en de resultaten ervan allemaal laat zien. Geen slappe verhaaltjes meer over voorzichtig interpreteren, maar laten zien hoe sterk je resultaat wordt beïnvloed door analytische keuzes die je maakte en waarvan je er meestal slechts één in een publicatie ziet. Dat is helaas nog te vaak dat resultaat dat de carrière van een supervisor of de beleidsbeslissing van een subsidiepartner het beste dient.

In multiverse-analyses loopt het aantal mogelijke analyses door het opstapelingseffect van de beslissingen al snel in de duizenden en wordt het weergeven ervan een uitdaging op zichzelf. Ga maar na, als we in het onderzoek over kettingroken 10 mogelijke verstorende variabelen hebben, zijn er al 1024 regressie-analysemodellen (2^10) denkbaar om voor die verstoringen te corrigeren. Dan zijn allerlei beslissingen

over sub-smaken van zo'n model nog niet eens meegenomen. Ook in niet-statistisch onderzoek moeten vele beslissingen worden gemaakt in de analyse. Zou multiverse analyse daar ook uitvoerbaar en presenteerbaar zijn? Het aantal publicaties met dit multiverse-analyses neemt de laatste 5 jaar gestaag toe (zie Figuur).



Figuur 2. Het aantal hits in Google scholar in 2020 t/m 2024 voor "multiverse analysis" OR "vibration of effect" OR "vibration of effects" OR "specification curve analysis"

En wie garandeert de vertrouwenswaardigheid van een multiverse-analyse? Preregistratie, open data en open code blijven de hoekstenen van transparantie. Maar het schrijven van de paragraaf over beperkingen van een multiverse-analyse blijft natuurlijk uitdagend. En zo reisden we in deze blog van afnemend vertrouwen in onze samenleving naar het belang van vertrouwenwekkend handelen en zagen dat er tools zijn om verder te komen dan slappe disclaimers in de discussiedelen van wetenschappelijke artikelen. De oude (2018) en de nieuwe Nederlandse gedragscode (2026) wetenschappelijke integriteit zegt tegen onderzoekers: "Wees in publieke communicatie eerlijk en helder over de beperkingen van het onderzoek (..)." Je begrijpt nu dat dat veel verder moet gaan dan het uitrekenen van een 95%-betrouwbaarheidsinterval rond één getal uit meer dan 1000 redelijkerwijs te verdedigen getallen.

Gerben ter Riet

1 oktober 2025

Met dank aan Heleen Wellner voor uitstekende suggesties en Figuur 1.