Steekproeven voor generalisatie – een belangrijke stap, maar we zijn er nog niet

Judith Schoonenboom[*]

Adri Smaling heeft een mooi en belangrijk artikel geschreven. Voordat ik inga op de resultaten van het artikel, wil ik eerst iets zeggen over de aanleiding, zoals die geschetst wordt in de eerste alinea. Deze bevat naar mijn mening een aantal misverstanden (waarbij ik niet wil zeggen dat de auteur zelf deze misvattingen huldigt), die ik graag uit de weg zou ruimen.

De eerste zin van het stuk, 'Kwalitatief onderzoek is van oudsher niet of minder gericht op generalisatie van de onderzoeksconclusies', roept bij mij direct vragen op als: 'Waarom niet? Minder dan wat?' Als wetenschappelijk product hebben de door kwalitatief onderzoek verkregen 'gedetailleerde kennis en inzicht, verdieping ervan en aandacht voor de gelaagdheid van ervaren betekenissen' toch ook tot doel bruikbaar te zijn, in ieder geval potentieel, in andere gevallen? Vanuit dat perspectief zou ik zeggen dat vrijwel alle onderzoek, kwalitatief of kwantitatief, gericht is op generalisatie, en dat verschillen gelegen zijn in de manier waarop er wordt gegeneraliseerd en waar de generalisatie uit bestaat, eerder dan in de gerichtheid op generalisatie.

Dit is een belangrijk punt, dat niet uit het oog verloren dient te worden. Ik herinner me goed dat ik als methodologisch adviseur een keer om advies werd gevraagd door uitvoerders van kwantitatief interventieonderzoek, die zich afvroegen of ze de uitkomsten van een effectmeting wel dienden te toetsen op significantie, omdat ze geen generaliseerbaarheid buiten de onderzochte groep (een schoolklas) nastreefden. Mijn reactie was dat ze de significantietoets wél dienden uit te voeren. Als je als (in dit geval onderwijskundig) onderzoeker een interventie of een module evalueert, doe je dat per definitie met het idee in je achterhoofd dat de resultaten daarvan potentieel iets zeggen over andere groepen, bijvoorbeeld de groep leerlingen die het jaar daarop hetzelfde onderwijs zal volgen. Het enige onderzoek dat ik kan bedenken dat niet gericht is op generalisatie is de census en aanverwante vormen: een (in het geval van de census: demografische) beschrijving van een situatie (in het geval van de census: de bevolking) op enig moment, met als enige doel deze in kaart te brengen. Op het moment dat je die in kaart brengt, doe je dat niet met het doel iets te kunnen zeggen over de toestand van volgend jaar, die immers geheel anders kan zijn.

Met de zin 'Kwalitatief onderzoek zou kleinschalig onderzoek zijn' wordt in ieder geval de indruk gewekt dat het kleinschalige karakter van kwalitatief onderzoek een reden zou zijn waarom generaliseerbaarheid in kwalitatief onderzoek niet aan de orde is. Achter deze indruk schuilen twee misverstanden. Ten eerste heeft (sta-

* Dr. Judith Schoonenboom is verbonden aan de opleiding Teaching and Learning in Higher Education van de vakgroep Onderwijswetenschappen en Theoretische Pedagogiek van de Vrije Universiteit Amsterdam. E-mail: judith.schoonenboom@vu.nl.

tistische) generaliseerbaarheid, anders dan wel wordt gedacht, niet te maken met het meest typerende van kleinschalig onderzoek, namelijk het (geringe) aantal participanten. Generaliseerbaarheid heeft te maken met het aantal observaties. Daarom vind ik de typering 'N=1'-onderzoek, die sommige kwantitatieve onderzoekers hanteren voor kwalitatieve studies, ook zo ongelukkig. Alsof onderzoek kan bestaan uit één observatie! Dat is precies het tegenovergestelde van de rijkdom aan ervaringen, processen en samenhangen die in kwalitatief onderzoek wordt blootgelegd. In onderzoek is het mogelijk, en in kwantitatief onderzoek ook vaak het geval, dat één individu gelijkstaat aan één observatie (bijvoorbeeld wanneer ieder individu steeds één keer dezelfde vraag beantwoordt). Maar er kunnen ook meerdere observaties per individu zijn. Denk daarbij aan herhaalde metingen bij hetzelfde individu, of denk aan vergelijking van hetzelfde construct in verschillende contexten (bijvoorbeeld het rapportcijfer op rekenen, het rapportcijfer op taal). In dergelijke gevallen zijn de observaties niet onafhankelijk van elkaar, maar in een statistische analyse kan daar heel goed rekening mee worden gehouden. Het rekenkundige aantal observaties is in zo'n geval weliswaar niet zo hoog als het aantal participanten maal het aantal metingen van hetzelfde construct, maar kan nog altijd vele malen hoger zijn dan het aantal participanten. Dat dit ook geldt voor kwalitatief onderzoek behoeft geen toelichting.
Dit impliceert dat, hoewel het aantal participanten in kwalitatief onderzoek doorgaans laag is, het aantal observaties dat in het geheel niet hoeft te zijn. Voor zover generaliseerbaarheid afhangt van de omvang van de steekproef, is het geringe aantal participanten in kwalitatief onderzoek slechts een bezwaar voor die eigenschappen waar individuen in het onderzoek slechts één keer op scoren. Dit geldt bijvoorbeeld voor achtergrondkenmerken als leeftijd en geslacht.
Een tweede misverstand is de suggestie dat de kern van statistische generalisatie bestaat uit het hebben van voldoende observaties. Ook dat is onjuist. Voldoende observaties vormen slechts een voorwaarde voor statistische generalisatie. De kern van statistische generalisatie is, zoals verderop in het artikel ook door Smaling betoogd, representativiteit: omdat de onderzochte groep representatief geacht wordt te zijn voor een bepaalde populatie, mag je de resultaten bij de onderzochte groep generaliseren naar die populatie.
Tot slot is de suggestie in de eerste alinea dat kwantitatief onderzoek sterker gericht zou zijn op generalisatie dan kwalitatief onderzoek, onjuist. In kwantitatief onderzoek mag uitsluitend statistisch gegeneraliseerd worden wanneer de onderzochte groep random (volgens een van de door Smaling genoemde methoden) getrokken is uit de populatie waarin men is geïnteresseerd; een eis waaraan naar een schatting van Tony Onwuegbuzie (p.c.) slechts zo'n 5 procent van het kwantitatieve onderzoek voldoet. Met andere woorden: zo'n 95 procent van het kwantitatieve onderzoek is niet statistisch generaliseerbaar. Ook kwantitatief onderzoek is dus doorgaans niet gericht op statistische generaliseerbaarheid.

Ik wil vier kanttekeningen plaatsen bij de in het artikel gepresenteerde resultaten. Een eerste, en meest belangrijke, kanttekening wordt zichtbaar op het moment dat ik probeer het geheel voor mezelf samen te vatten:

- Generaliseren op basis van representativiteit (statistische generalisatie of variatiedekkende generalisatie) vereist een representatieve steekproef (aselect, volledig systematisch of quota).
- Steekproeftrekking kan op twee manieren bijdragen aan theoretische generalisatie. Ten eerste door participanten te kiezen op basis van relevante inherente kenmerken (typisch geval of kritisch geval). Ten tweede door via herhaling (replicatief, iteratief of theoretisch) de interne validiteit van het onderzoek op een aanvaardbaar peil te brengen. Op het moment van saturatie bereikt men twee dingen tegelijk: doordat toevoeging van leden uit de populatie niet langer leidt tot wijziging van de theorie bereikt men zowel theoretische saturatie (interne validiteit) als generalisatie (want een nieuwe participant toevoegen levert geen tegenstrijdige informatie meer op).
- Generalisering door overdracht gebeurt op een *case-by-case* basis. Daarom is representatieve steekproeftrekking hiervoor niet geschikt. Steekproeftrekking op basis van relevante gevalskenmerken (typisch of kritisch) is dat wél. Een steekproeftrekking door voortdurende toevoeging van nieuwe gevallen is ook geschikt.
  NB Ik moet bekennen dat ik er niet helemaal zeker van ben dat ik dit laatste punt goed heb begrepen.

Deze samenvatting van mijn eigen begrip verschilt op een aantal punten van de matrix en van wat daarover door Smaling gezegd wordt. Ten eerste zijn in mijn begrip statistische generalisatie en variatiedekkende generalisatie hetzelfde, namelijk generalisatie op basis van representativiteit. Je kunt de twee onderscheiden, maar de vraag is waarom je dat zou doen. Ik zie de matrix als een belangrijk hulpmiddel voor kwalitatieve onderzoekers, die op basis van de door hen nagestreefde vorm van generalisatie kunnen kiezen voor een bepaalde vorm van steekproeftrekking. Voor deze onderzoekers is, zoals Smaling ook aangeeft, de aselecte steekproeftrekking slechts van belang vanwege de representativiteit; er wordt niet statistisch gegeneraliseerd. Waarom zou je dan statistische generalisatie opnemen als specifieke vorm van generalisatie via representativiteit?
Met betrekking tot de steekproeftrekking zou ik de vormen gericht op generalisatie via representativiteit samen willen nemen, en ze daarbij willen zien als subtypen. Ik doe dat deels om, op milde wijze, te provoceren. Ik vraag me namelijk af of de nadruk die er vanuit kwantitatieve hoek wordt gelegd op het verschil tussen aselecte en systematische steekproeftrekking terecht is. De redenering is dat alleen door vormen van aselecte steekproeftrekking bij een voldoende grote steekproef statistische generalisatie kan worden verkregen, omdat het bij een systematische steekproef altijd denkbaar blijft dat er niet-gecontroleerde verschillen tussen groepen (typisch een experimentele en een controlegroep) blijven bestaan. Er dient, aldus de kwantitatieve canon, een match te zijn tussen de vorm steekproeftrekking, de gehanteerde analyse en de getrokken conclusies. Die veronderstelling is echter onjuist. Dit laat zich illustreren aan de hand van een vergelijkbaar vereiste in kwantitatief onderzoek, namelijk dat parametrische toetsen intervaldata vereisen, en dat men daarom op data verkregen met Likertschalen (helemaal mee oneens … helemaal mee eens) geen parametrische toetsing (zoals

het uitrekenen van een gemiddelde) mag verrichten. De juiste vraag is echter niet of datatype en analyse op elkaar passen. De juiste vraag is: hoe schadelijk is het voor je conclusies als je parametrische toetsing loslaat op Likertschaaldata? Het antwoord, gegeven door enkele gerenommeerde statistici in de jaren vijftig en door Geoff Norman in 2010 op basis van simulaties, is: dat is helemaal niet erg. Naar analogie is de vraag met betrekking tot steekproeftrekking en statistische generaliseerbaarheid: hoe erg is het voor de statistische generaliseerbaarheid als de (voldoende grote) steekproef niet random is getrokken maar systematisch? Ik ken het antwoord op deze vraag niet, maar mogelijk is het al wel gegeven.
Op een vergelijkbare manier lijkt het mij verstandig om ook een aantal andere categorieën uit de matrix samen te brengen, en te werken met subgroepen. Ik ga daar nu niet verder op in.
Mijn tweede kanttekening betreft de plaats van de matrix in relatie tot andere doelen van steekproeftrekking. Smaling heeft zich in dit stuk terecht beperkt tot één, vaak onderbelichte, functie van steekproeftrekking: de generalisatie. Maar steekproeftrekking dient nog een ander doel, dat bij Smaling weliswaar impliciet aan de orde komt, maar niet expliciet als doel van steekproeftrekking wordt genoemd: het optimaliseren van theorievorming; in kwantitatief jargon: de interne validiteit. Ook dat doel dient in kwalitatief onderzoek te worden bereikt, en de vraag dringt zich dan ook op hoe een onderzoeker aan deze beide eisen van interne validiteit en generaliseerbaarheid kan voldoen, en of, en zo ja, hoe deze tegenover elkaar afgewogen dienen te worden. In dit verband is de theoretische steekproeftrekking interessant, omdat daarin beide eisen samenvallen: op het moment dat theoretische saturatie is bereikt (doel: interne validiteit), wordt ook het doel van de theoretische generalisatie bereikt (een nieuwe participant toevoegen levert geen wijziging in de theorie meer op).
Als derde kanttekening mis ik de vorm van generaliseerbaarheid die door Shadish, Cook en Campbell (2002: 83) is aangeduid als generalisatie over 'variaties tussen personen, omgevingen, behandeling en uitkomsten die *in* het experiment waren', in plaats van erbuiten.
Tot slot, als vierde kanttekening, mis ik een discussie over het bereik van de generalisatie. Dit is in veel kwantitatief onderzoek een probleem, maar speelt net zo goed bij kwalitatief onderzoek. Zo wordt het overgrote deel van het psychologisch onderzoek uitgevoerd bij niet random gekozen selecties van psychologiestudenten. De resultaten daarvan zijn, al wordt dit doorgaans verzwegen, niet statistisch generaliseerbaar (maar wel theoretisch). Het bereik van de generalisatie is echter beperkt tot bijvoorbeeld psychologiestudenten aan Amerikaanse universiteiten, terwijl doorgaans een veel breder bereik wordt nagestreefd.

Ik kijk uit naar het vervolgartikel.

Literatuur

Norman, G. (2010). Likert scales, levels of measurement and the 'laws' of statistics. *Advances in Health Sciences Education, 15*(5), 625-632. doi:10.1007/s10459-010-9222-y.

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