Kroniek

Verslaggeving over empirisch onderwijskundig onderzoek ten onzent

0. Onderwijskundige monografieën, leerboeken en handboeken – vooral Amerikaanse, maar in navolging hiervan ook ten onzent – spreiden een geweldig aangroeiende hoeveelheid van zogenaamd empirisch onderzoek ten toon, waarbij 'empirisch' meestal te verstaan is in de zin van 'statistisch vergaard en bewerkt'. De stijl varieert tussen twee uitersten: van het alleen maar aanhalen van problemen die elders bestudeerd zijn, tot beweringen omtrent hetgeen met zekerheid bewezen of weerlegd zou zijn. De wegwijzers naar het bewijsmateriaal leiden naar een uitgebreide bibliografie aan het eind, en dat is dan, wat de lezer betreft, ook het eind van de weg. Er wordt naar omvangrijke werken verwezen, zonder de geciteerde bladzijden of paragrafen nader te specificeren; of de aangehaalde artikelen zijn resumé's van ongepubliceerde omvangrijke rapporten, die voor buitenstaanders practisch ontoegankelijk zijn, wanneer ze dan na verloop van vijf of tien jaren in het geheel nog mochten bestaan.

Als redacteur en adviseur van onderwijskundige tijdschriften en instellingen met schade en schande wijs geworden, heb ik vanaf zeker ogenblik bij elk manuscript dat ik in handen kreeg het bewijsmateriaal opgevraagd – soms een langdurige procedure als men echt tot de bronnen terug wil. De resultaten waren ontstellend: ik schat dat minder dan één op de tien manuscripten de toets van nader onderzoek kon doorstaan. De tekortkomingen varieerden van administratieve vergissingen (zoals het abusievelijk verwisselen van onderdelen van experimentele en 'control' groep of het abusievelijk inconsistent nummeren van toetsen) tot ondeugdelijke of ondeugdelijk geëvalueerde toetsinstrumenten toe, van ondeugdelijke steekproeven tot onjuiste statistische procedures.

Zou de steekproef van manuscripten die ik te zien kreeg, representatief zijn voor de betrouwbaarheid van empirisch onderwijskundig onderzoek? Ik vatte het plan op dit na te gaan. Ik zou uit een groot handboek van onderwijskundig onderzoek een aselecte steekproef van honderd verwijzingen naar zogenaamd empirisch onderzoek trekken en deze bronnen nader bekijken om vast te stellen of de citaten en daaruit getrokken conclusies terecht waren. Ik kwam niet ver met deze onderneming. Het bleek spoedig dat soms de citaten te onnauwkeurig waren en dat in vele andere gevallen op de geciteerde plaats verwezen werd naar practisch ontoegankelijke literatuur.

Het toeval wilde dat ik, een werk van B. S. Bloom[1] lezende, op beweringen stuitte die bij me twijfels opriepen. Deze beweringen – en andere – bleken gebaseerd te zijn op ongepubliceerde dissertaties van leerlingen van de auteur. Ik slaagde erin me van een deel ervan microfilms te verschaffen. Ik heb de dissertaties geanalyseerd en hun uitkomsten met de citaten bij Bloom vergeleken. Van mijn bevindingen heb ik uitvoerig[2] alsmede in een verkorte versie[3] verslag uitgebracht. Een der onderzochte dissertaties bleek geheel ondeugdelijk te zijn; sommige andere vertoonden minder ernstige tekortkomingen. Het voornaamste resultaat was echter dat Blooms rapportage op vele punten onjuist was en dat de door Bloom geciteerde bronnen zijn conclusies in genen dele rechtvaardigden.

Met deze analyse is mijn vraag naar de betrouwbaarheid van zogenaamd empirisch onderzoek uiteraard niet beantwoord. Mijn steekproef was zeker niet aselect – het waren allemaal proefschriften van één universiteit.

1. In de loop der jaren heb ik me ook – ambtshalve en anderszins – met in het Nederlands gepubliceerd onderzoek bezig gehouden. Deze analyses zijn, voorzover in het geheel, slechts in kleine kring – de auteurs inbegrepen – verspreid. Een der laatste gevallen was een bijdrage van K. Hesselink tot een bundel onder redactie van E. Warries[4]. Mijn gewoonte getrouw heb ik mijn bevindingen aan de auteur van het artikel en de redacteur van de bundel medegedeeld. Hun reactie was de aanleiding voor het hier volgende uittreksel uit mijn analyse, dat zich vrijwel geheel beperkt tot wat in de titel genoemd wordt de wijze van verslaggeving.

2. Hesselink bericht op 15 bladzijden over een experiment, waarvoor dat stuk de enige informatiebron is. Door omstandigheden beschik ik over meer informatie, in bladzijden gemeten 50-100 keer zo veel. Dit lijkt veel, maar is in feite een ontoereikende fractie van het totale materiaal. Het beslaat echter al hetgeen in 1976 (nog) daaromtrent verkrijgbaar was, alsmede aanvullingen van latere datum. Het materiaal bevat projectaanvragen, voortgangsrapportages,

proeflessen, summatieve en retentie-toetsen, scorelijsten, instrumenten voor 'affectief' onderzoek en het Eindverslag van 1978. Tussen het primaire materiaal en de verslagen over het experiment bestaan discrepanties. Na een half jaar vergeefse pogingen ze op te helderen, ga ik tot publicatie over.

3. Nadat men bij het P.C. Enschede – op zijn minst sinds 1969 – ervaringen met rekenonderwijs – vooral in niveaugroepen – had opgedaan, voelde men aldaar de behoefte een eigen methode – Gedifferentieerd Rekenen – te ontwikkelen en deze ontwikkeling in de onderzoeksfeer te trekken. Door E. Warries, inmiddels hoogleraar aan de T. H. Twente, werd een SVO-project (0289) aangevraagd en verkregen, aanvankelijk voor 1975, dan met één jaar verlengd. Het was de bedoeling van het onderzoek, Mastery Learning met andere strategieën te vergelijken. Hierbij zouden drie rekenmethoden betrokken worden: Gedifferentieerd Rekenen, door het P.C. Enschede ontwikkeld, als voorbeeld voor Mastery Learning, en twee commercieel verkrijgbare methoden als 'control': 'Niveaucursus Rekenen' en 'Nieuw Rekenen voor het Basisonderwijs'. Volgens het projectplan zouden deze drie methoden geanalyseerd worden op hun minimum doelstellingen, de doelstellingen van hun verrijkingspakket en de aan het rekenonderwijs bestede tijdsduur. Er zou summatief getoetst worden in hoeverre bij de afzonderlijke methoden de respectievelijke doelstellingen bereikt werden en in welke mate de resultaten spreidden. In de affectieve sfeer zouden de opinies van leerlingen en onderwijzers omtrent het onderwijs volgens de afzonderlijke methoden worden gepeild.

4. In het voor mij beschikbare materiaal ontbreken lijsten van minimum doelstellingen en doelstellingen voor verrijkingsstof voor de drie methoden. Ik beschik over geen materiaal waaruit blijkt dat deze ooit zijn vastgesteld. Bij mijn weten was al hetgeen aan de cognitieve kant is geschied: de leerlingen op het eind van het eerste leerjaar een summatieve toets en na de vakantie een retentietoets over stipsommen (alsmede verrijkingstoetsen) af te nemen, bestaande uit 40 items (respectievelijk 24 punten opbrengende). Alle toetsen, alsmede de enquêtelijsten werden door medewerkers van het P.C. Enschede vastgesteld, afgenomen en geëvalueerd. Het onderzoek is niet vervolgd, het Eindverslag niet gepubliceerd.

5. Bij de drie-methoden-opzet van het project bestond het gevaar, dat de variabele 'onderwijsstrategie', die men wilde evalueren, vertroebeld werd door de variabele 'onderwijsleerstof'. Dit gevaar zou ontweken worden door elke methode op zijn eigen merites te toetsen. Deze voorzichtigheid is bij de uitvoering echter niet betracht. De drie groepen werden op één onderwerp getoetst, met één instrument. Om de onderwijsstrategieën statistisch te kunnen vergelijken moet men er zeker van zijn dat de factor 'onderwijsleerstof' de resultaten niet had beïnvloed, althans in staat zijn deze factor te elimineren. De factor bestaat uit twee componenten: het rekenonderwijs voorafgaande aan het onderwerp stipsommen en het onderwijs in stipsommen. Over de eerste component had men zich inlichtingen kunnen verschaffen door middel van een toets op rekenvaardigheden in het algemeen. Bij mijn weten is zo'n ingangstoets niet afgenomen. De vraag omtrent de tweede component werd door de projectleider in oktober 1976 positief beantwoord: 'Voor het onderdeel puntsommen is ervoor gezorgd dat het didactisch uitgangspunt in alle drie de groepen gelijk was (leertheorie Gal'perin).' In correspondentie – februari 1980 – werd deze uitspraak teruggebracht tot die dat de gelijkwaardigheid wel was nagestreefd, maar dat niet vaststaat of die poging gelukt is. Zowel in het Eindverslag als in het geciteerde stuk ontbreekt elke verwijzing naar het voor de vergelijking van de strategieën cruciale punt van de 'gelijke start'. Dit lijkt me een ernstig tekort in de verslaggeving.

6. In het materiaal waarover ik beschik, ontbreken gegevens waaruit blijkt of en op welke wijze de methode 'Gedifferentieerd Rekenen' *in feite* volgens de regels van Mastery Learning is onderwezen. In februari 1980 deelde de toenmalige projectleider mij mede dat harde gegevens hieromtrent niet beschikbaar zijn. Ik kan de redenen waarom hij de feitelijke gegevens niet publiek toegankelijk maakte, geheel respecteren. Desalniettemin lijkt het mij een gebrek in de verslaggeving dat de gegronde twijfels omtrent het Mastery karakter van 'Gedifferentieerd Rekenen' niet kenbaar zijn gemaakt.

7. Het in het Eindverslag gepresenteerde statistische materiaal is verdeeld over het cognitieve en het affectieve aspect. De rapportage over het affectieve aspect beslaat het overgrote deel van het Eindverslag: correlaties, variantie-analyses en multiple regressies – een nachtmerrie van cijfers. Wie het Eindverslag meer dan oppervlakkig bestudeert, zal opmerken dat het verslag over het cognitieve aspect op een andere leerlingenpopulatie slaat dan dat over het affectieve aspect. In het eerste geval is er sprake van 15 klassen, in het tweede geval van 18. In de scorelijsten, waar ik over beschik, zijn het ook 18 klassen. Deze discrepantie wordt in de voorlaatste alinea van hoofdstuk 1 van het Eindverslag opgehelderd. De definitieve tabellen en grafieken waarin de deelnemende klassen qua score en mastery werden vergeleken, zijn tot stand gekomen nadat men drie van die klassen had geschrapt: 'Deze klassen bleken bijzonder significant te verschillen met de andere klassen. We besloten dat dit moest voortkomen uit het niet volgen van de schriftelijk gegeven instructies.' Merkwaardigerwijs zijn de geschrapte klassen wél aangehouden bij de bewerking van de affectieve data.

Het is niet eenvoudig na te gaan *welke* drie klassen geschrapt werden. De coderingen in het cognitieve en in het affectieve deel van het Eindverslag kloppen niet met elkaar; de scorelijsten waar ik over beschik, geven weer een andere nummering. Na veel passen en meten ben ik tot de conclusie gekomen dat de drie afgevoerde klassen zijn: – de twee laagst scorende uit de groep Mastery Learning; – de hoogst scorende uit de groep Niveaucursus Rekenen. De toenmalige projectleider was – desgevraagd – in februari 1980 niet in staat deze conclusie te bevestigen. Tijdgebrek belette hem tot nu toe na te gaan welke klassen geschrapt zijn en hoe de diverse coderingen met elkaar gerelateerd zijn. Ik

kan trouwens de motieven van de versluiering van de codering respecteren: de betrokken scholen niet te laten weten dat ze van onregelmatigheden verdacht werden.

In februari 1980 deelde de toenmalige projectleider mij ook mede dat in tegenstelling met het in het Eindverslag beweerde bij het schrappen *niet* van significante verschillen sprake was. Mijns inziens hadden bij een behoorlijke verslaggeving in het Eindverslag naast de geschoonde ook de 'vuile' gegevens moeten worden verstrekt en had in het geciteerde artikel het feit van de schoning moeten worden vermeld.

8. Volgens de grafieken in het Eindverslag en in het geciteerde artikel werd 100% mastery behaald bij de summatieve toetsen door plm. 30%, 25% en 3% van de leerlingen respectievelijk in de Mastery groep, de Niveaurekengroep en de Nieuw-Rekenen-groep. Bij de retentietoetsen waren de percentages 40%, 24% en 12%.

100% Mastery zou betekenen: 40 van de 40 sommen goed. Wie iets van het onderwijs afweet, zal bij het zien van deze enorme percentages meesterlijke rekenaars (op het eind van het eerste en in het begin van het tweede leerjaar) zijn ogen uitwrijven. Volgens mijn scorelijsten liggen de percentages van 100% mastery tussen de 1% en 3%. Hoe zijn die grafieken te rijmen met de scorelijsten?

Het is duidelijk dat de mastery van de grafieken niet is berekend volgens de voor de hand liggende procedure van 'zoveel percent van de 40 sommen goed'. Ik heb van alles geprobeerd om de procedure te achterhalen: afrondprocedures zoals door Warries in hetzelfde boek uiteengezet, weglaten van wat ik voor de drie 'zondige' klassen aanzag, samenvatten van de items in acht sets van vier en een van acht, met diverse criteria voor slagen in een set, en dit al met elkaar gecombineerd. Het mocht niet baten. Tenslotte heb ik me tot de toenmalige projectleider gewend met herhaalde verzoeken om inlichting. Ook dit mocht niets baten.

Ik beklemtoon met alle nadruk dat ik de projectleiding niet van knoeien met de gegevens verdenk. Het is duidelijk dat het criterium 'mastery' een herwaardering heeft ondergaan, die redelijk gemotiveerd zal zijn, maar kennelijk is de formule volgens welke de mastery werd berekend, thans niet meer te achterhalen.

Het is in elk geval een zeer ernstig gebrek in de verslaggeving dat geen inlichtingen worden verstrekt over de wijze waarop mastery werd berekend.

9. Ik heb me hier tot tekortkomingen in de verslaggeving beperkt. Aan het onderzoek zelf kleven fouten. Ik kan tenslotte niet nalaten twee bijzonder belangrijke op te noemen.

Ten eerste, bij de correlatie-, variantie- en regressie-analyses is de score in plaats van de mastery als afhankelijke variabele gekozen.

Ten tweede: één blik op de ('vuile' of geschoonde) scorelijsten laat zien dat de diverse deelnemende klassen onderling enorm verschilden, wat hun prestaties aangaat: klassen waar slechts 10% van de leerlingen 80% mastery behalen naast klassen waar dit voor 75% het geval is. Het verschil tussen de drie *groepen* (beantwoordende aan de drie *methoden*) lijkt in het niet te vallen, vergeleken bij het verschil tussen de *klassen*. Of dit inderdaad het geval is, kan met een klassieke significantie toets worden vastgesteld. Er zijn geen aanwijzingen dat dit overwogen is. Eigenlijk had het al in een vooronderzoek kunnen en moeten geschieden. Er zijn er die menen dat het er voor studieprestaties nauwelijks toe doet welke methode wordt gebruikt. Wat echt telt, zouden het schoolklimaat en de onderwijzer zijn. Kiest men voor zulk een onderzoek klassen op goed geluk, dan is de kans groot dat het statistisch misloopt. Bij statistisch onderzoek, zoals hier beoogd, pleegt men de populaties door 'matching' te constitueren. Kennelijk was dit niet haalbaar. Maar dan was het falen ook ingebakken.

H. Freudenthal

Noten

1. B. S. Bloom. *Human Characteristics and School Learning*. New York: McGraw-Hill 1976, XII + 284 p. Speciaal blz. 18, 23, 25, 55-56.
2. H. Freudenthal. Ways to Report on Empirical Research in Education. *Educational Studies in Mathematics*, 1979, *10*, 275-303.
3. De waarde van resumerende en tweedehands informatie. *Pedagogische Studiën*, 1979, *56*, 323326.
4. E. Warries, e.a. *Beheersingsleren – een leerstrategie.* Groningen: Wolters-Noordhoff, 1979. Bijdrage 5: K. Hesselink. *Beheersingsleren in de lagere school, een opzet bij het rekenonderwijs*, blz. 42-56.

Manuscript aanvaard 21-10-'80