



Beeld / Herman Wouters

Mensen zijn geen getallen. Helemaal waar. Maar toch kan menselijk gedrag voor een deel prima in getallen worden gevat, meent statisticus Cees Glas. Het toetsen van gangbare kennis, meningen en overtuigingen aan feitenmateriaal kan soms onverwachte inzichten opleveren. Dat geldt ook voor het beoordelen van de kwaliteit van onderwijs, en de prestaties van leerlingen. Niet altijd leuk, maar wel vaak nuttig.

Door Cees Glas

Meten met mate

De meetbare waarde van onderwijs

De tegenstelling tussen de koele, op harde empirische gegevens gebaseerde oriëntatie en de naturalistische oriëntatie op sociale wetenschap is al behoorlijk oud. In de sociologie is ze al terug te vinden in de tegenstelling tussen positivist Emile Durkheim (1858–1917) en zijn tegenvoeter Max Weber (1864–1924). Inleidende colleges methodenleer en statistiek beginnen vaak met een verhandeling over een belangrijke bijdrage van Durkheim, de zogenaamde ‘elaboratietechniek’. Deze techniek maakt duidelijk dat observeerbare samenhangen tussen verschijnselen (correlaties) nog niet betekenen dat er ook sprake is van een oorzakelijk verband. Durkheim liet dit zien aan de hand van het verband tussen de huwelijkse staat en het plegen van zelfmoord. De correlatie was positief en velen trokken daaruit de conclusie dat de zorg voor een gezin de kans op zelfmoord vergrootte. Men schreef het fenomeen toe aan de slechte sociaaleconomische omstandigheden uit die dagen. De elaboratietechniek is het controleren van achtergrondvariabelen die een mogelijke alternatieve verklaring voor een samenhang kunnen opleveren. Durkheim gebruikte leeftijdscategorieën en liet zien dat binnen de leeftijdscategorieën het verband tussen een huwelijk en zelfmoord niet bestond. Het was niet de druk van de sociaaleconomische omstandigheden, maar veeleer het ongebonden zijn en de vereenzaming die tot zelfmoord leidden. Inmiddels is Durkheims verklaring gemeengoed geworden. Zijn techniek was het begin van de moderne toepassing van causale statistische modellen voor het blootleggen van sociale en psychologische mechanismen.

Een tweede voorbeeld van de manier waarop zorgvuldige statistische analyses psychologische mechanismen kunnen verklaren komt uit Nederland. Veel ouders hebben het vaak over verschillen tussen het oudste en het jongere kind. Het gaat daarbij om anekdotische verhalen die moeilijk hard te maken lijken. Maar er bestaat cijfermateriaal. Eind zestiger jaren werd er een grootschalig onderzoeksproject uitgevoerd om de effecten van ondervoeding op de cognitieve vaardigheden van kinderen die geboren waren rond het einde van De Tweede Wereldoorlog te bestuderen. Als onderdeel van het onderzoek werd een IQ-test, de Raven test, afgenomen bij

350.000 jongeren op het tijdstip dat ze 19 jaar oud werden. De onderzoekers vonden vier opmerkelijke fenomenen. In de eerste plaats scoorden de kinderen uit kleine gezinnen gemiddeld hoger dan kinderen uit grote gezinnen. Dit verschil was

‘Mensen zijn geen getallen. Maar het toetsen van gangbare kennis, meningen en overtuigingen aan feiten kan soms onverwachte inzichten opleveren, of gangbare, vaak dierbare kennis ondergraven’

consistent voor iedere gezinsgrootte tussen twee tot en met negen kinderen. In de tweede plaats bleek dit verschil niet te gelden voor gezinnen met één kind. Daar lag het gemiddelde tussen dat van gezinnen met drie en vier kinderen. In de derde plaats was er een significante statistische relatie tussen de geboortevolgorde en de intelligentie: het oudste kind scoorde gemiddeld het hoogste en daarna nam het gemiddelde steeds in dezelfde mate af. Tenslotte vond men een extra afname bij het laatste kind. Lilian Belmont en Francis Marolla, beiden van de Columbia Universiteit in de Verenigde Staten publiceerden hun resultaten in het wetenschappelijk toptijdschrift *Science* in 1973. Ze lieten zien dat sociaaleconomische verschillen tussen gezinnen met een verschillend kindertal de resultaten niet konden verklaren.

De sociaal psychologen Robert Zajonc en Greg Markus kwamen in 1975 met een model dat de resultaten wel verklaarde. Zij veronderstelden dat ieder gezin een zogenaamd ‘intellectueel niveau’ heeft, dat bepaald wordt door alle leden,

Metten zet ons op het juiste spoor

Iedere dag stappen velen van ons achteloos met een gerust hart in de trein: het ergste wat je kan overkomen is stilstand. Ergens tussen de weilanden. Zomaar, om onduidelijke redenen. Een treinongeluk? Dat zal toch wel niet? Zolang de machinist zijn blik op scherp heeft, onverwachte obstakels uitblijven en de trein in het juiste spoor blijft gaat alles goed: volkomen veilig, die trein.

Maar vergis u niet. Hebt u zich ooit gerealiseerd, terwijl u zich 's ochtends vroeg nog maar halfwakker naar uw werkplek begeeft, hoe nauw het luistert een trein in het spoor te houden? 1435 mm is de afstand tussen twee spoorstaven op een recht stuk spoor. In 'een boog' (bocht) is dat 1450 mm. Een paar millimeter meer of minder en de trein ontspoord. Hetzelfde zal gebeuren wanneer de hoogteligging van de spoorstaven ten opzichte van elkaar een onjuiste afwijking vertoont. Zeker bij overgang van recht spoor naar 'een boog' kent dit nauwe toleranties. En wat te denken van de maatvoering van een wissel...

Toch denderen al die volle treinen over het spoor: iedere dag weer. Hebt u bij een overweg wel eens opgemerkt wat een krachten een langzame trein op het spoor uitoefent? Dat 'ie er niet uit vliegt is dankzij een proces van voortdurend meten. Alles, iedere meter spoor, iedere wissel, wordt gemeten, opnieuw gemeten en weer gemeten.

Vroeger liepen er dagelijks inspecteurs langs het spoor, met meetinstrumenten en matenkaartjes. Tegenwoordig worden geavanceerde meetreinen ingezet. Deze meten met lasers en fotografieren ieder stukje spoor. De vroegere inspecteurs zitten nu achter beeldscherm. Zij interpreteren de door de slimme software herkende veranderingen, waarna – meestal 's nachts – herstploegen op pad gaan. Zo worden dreigende overschrijdingen van veiligheidstoleranties voorkomen en kunt u morgenochtend met een gerust hart die trein weer in.

Ferry Visser
Salesmanager bij RailCom

dus ouders zowel als kinderen. De bijdrage van de ouders werd op 100 gesteld. De bijdrage van een pasgeboren baby werd op nul gesteld en daarna droeg het kind ieder jaar meer bij. De zo geconstrueerde variabele Intellectueel Niveau daalde met het aantal aanwezige kinderen en verklaarde de algemene neerwaartse tendensen van het IQ. Echter niet de extra dip van het enig kind en het laatste kind. De verklaring van Zajonc en Markus voor die twee verschijnselen was dat kinderen behalve van hun omgeving ook leerden door zelf als mentor voor jongere kinderen op te treden. Enige kinderen en laatste kinderen krijgen die kans echter nooit. Vervolgonderzoek naar geboortevolgorde toonde verder verschillen tussen kinderen aan met betrekking tot angst, agressie, sociale vaardigheden en avontuurlijkheid. Op zich is het interessant dat rigoureuus onderzoek onder het publiek levende vermoedens kan

‘Objectieve toetsen hebben een emancipatoir effect’

onderbouwen. Er moet echter wel een kritische kanttekening gemaakt worden. Het gaat om gemiddelde effecten. Door de zeer grote steekproef van 350.000 jongeren zijn de gevonden effecten zeer significant, dat wil zeggen, niet aan toeval toe te schrijven. Verder zijn de effecten bijzonder consistent. Maar de effecten zijn ook bijzonder klein. Het verschil tussen de eerstgeborene in een gezin met 2 kinderen en de laatstgeborene in een gezin met 9 kinderen was 5 IQ-punten. Het effect bestaat dus overduidelijk, maar het is klein en voor een individueel kind kan je er absoluut geen conclusie aan verbinden.

Metten in het onderwijs: de scholen

De vraag is natuurlijk of dit soort statistische analyse ons ook kan helpen bij het beoordelen van de kwaliteit van het onderwijs. Grofweg vinden er in het onderwijs twee soorten metingen plaats:

metingen gericht op het onderwijssysteem en metingen gericht op de leerling. Laten we eerst naar metingen gericht op het onderwijssysteem kijken. Veel landen volgen de ontwikkelingen van de output van het onderwijssysteem met peilingsonderzoek. In Nederland is dat het PPON, de Periodieke Peiling van het Onderwijs Niveau. Het gaat hierbij om een evaluatie op macroniveau gericht op de kwaliteit van de output van ons onderwijssysteem. Daarnaast bestaan er in Nederland onderzoeken naar de effectiviteit van het schoolstelsel die meer in de diepte gaan en de output proberen te verklaren, zoals de cohortstudies in het kader van COOL 5-18. Leerlingprestaties op het gebied van Nederlandse taal, rekenen en wiskunde worden in deze studies voornamelijk gebruikt als indicatoren voor beleidseffecten. Tot slot doet Nederland ook mee aan grootschalig internationaal vergelijkend onderzoek, zoals TIMMS (Trends in International Mathematics and Science Study), PIRLS (Progress in International Reading Literacy Study) en PISA (Program for International Student Achievement). Het gaat hierbij om drie vergelijkbare projecten waarin leerprestaties van kinderen in een groot aantal landen met elkaar vergeleken worden. Bij het laatstgenoemde onderzoek (PISA) worden kinderen uit een groot aantal landen in een driejaarlijkse cyclus getest op leesvaardigheid en wis- en natuurkunde. In 2009 deden er 65 landen en bijna een half miljoen leerlingen mee. De rangorde van de prestaties van de landen blijkt het meest tot de verbeelding van politici en de pers te spreken. Die rangorde is over de jaren heen opvallend stabiel: Finland scoort bijvoorbeeld altijd in de top-drie en Nederland rangeert altijd tussen de 9e en 11e plaats. Relatief doet Nederland het dus goed. Bijvoorbeeld veel beter dan de Verenigde Staten.

Veel interessanter is echter dat in het PISA-project ook heel veel achtergrondvariabelen van kinderen, ouders en scholen gemeten worden. De relatie tussen die achtergrondvariabelen en de leeropbrengsten kunnen inzicht geven in de werking van het schoolstelsel. Een van de interessantste relaties is die tussen de sociaaleconomische achtergrond van een leerling en de leerprestaties. Als we gelijke kansen serieus nemen,

zou die relatie verwaarloosbaar moeten zijn. Geen enkel land voldoet daaraan. Finland heeft een egalitair en niet-gesegregeerd schoolstelsel, waarin leerlingen pas laat hun definitieve keuze voor hun uiteindelijke opleiding maken. In Finland is de relatie tussen de sociaaleconomische achtergrond en de leerprestaties dan ook het zwakste. Nederland doet het daarin slechter. En wat nog opmerkelijker is, is dat het niet de individuele sociaaleconomische achtergrond van een leerling is die het verschil maakt, maar het schoolgemiddelde van de sociaaleconomische achtergrond van de leerlingen. Schijnbaar is er in Nederland dus sprake van een behoorlijke segregatie tussen de scholen. Maar wel weer veel minder dan bijvoorbeeld in Groot-Brittannië of de Verenigde Staten. Toch moeten we ondanks deze resultaten het belang van schoolkeuze niet overdrijven. Het percentage variantie van leerlingprestaties dat in Nederland door de school verklaard wordt, is ongeveer 15 procent. Ouders sloven zich bijvoorbeeld uit om voor hun kinderen een 'goede school' te zoeken. Maar voor het niveau van de leerprestaties maakt het uiteindelijk heel wat minder uit dan al die ijverige ouders denken.

‘Metten is een tamelijk bescheiden onderneming’

Metten in het onderwijs: de leerling

Bij overheden en ouders vindt heel wat wensdenken plaats. Overheden hopen dat hun onderwijsbeleid en hun onderwijsvernieuwingen zoden aan de dijk zetten. Onderwijskundig metten is nuttig om de effecten daarvan te toetsen. Bij ouders ligt het nog wat gevoeliger. Bij hen kan wensdenken zich vertalen in de overtuiging dat hun kind een aankomend hoogbegaafd genie is. Onderwijskundig metten kan in zo'n geval een pijnlijke confrontatie met de realiteit opleveren. Vaak richt de boosheid zich op de Cito-eindtoets en de centrale eindexamens. Dat is niet terecht.

Zo zijn objectieve toetsen zoals de Cito-eindtoets slechts een onderdeel in de toelating tot het vervolgonderwijs, naast docentoordelen en de opinie van de ouders. Die subjectieve oordelen zijn heel waardevol, maar uit onderzoek blijkt dat verwachtingen van docenten, leerlingen en ouders elkaar ook op een heel ongewenste manier kunnen beïnvloeden. Een objectief gegeven naast het docentenoordeel kan daarom geen kwaad. Ook verschillen tussen de centrale eindexamens en schoolonderzoeken laten zien dat het niet onverstandig is om bij belangrijke beslissingen externe metingen te betrekken. Daarnaast bieden die externe toetsen leerlingen waarvan je dat niet zou verwachten de kans te laten zien wat ze kunnen. In die zin hebben objectieve toetsen ook een emancipatoir effect.

Er is nog een argument waarom toetsing belangrijk is. Nederland heeft een grote vrijheid van onderwijs met een veelheid van schoolsoorten: openbaar, katholiek, protestants-christelijk, Montessori, Frenet, Jena Plan, Dalton, Vrije Scholen op Antroposofische basis, enzovoort. Om ervoor te zorgen dat leerlingen toch enigszins vergelijkbare eindniveaus halen, zijn er door de overheid referentieniveaus vastgesteld die vervolgens weer vertaald zijn in een verzameling toetsen die de overgang van het primair onderwijs naar het vervolgonderwijs begeleiden. Door vaker te toetsen gaan we meer van een momentopname naar het volgen van een ontwikkeling, diagnostiek, en eventuele tijdige bijsturing. Door het element van een momentopname minder te accentueren en toetsing meer in het onderwijs te integreren, wordt de druk op de leerling ook verder verminderd.

Ten slotte

Metten is weten, hoor je wel eens. In feite is dat niet waar. Als het goed is, gaan metingen altijd vergezeld van betrouwbaarheidsintervallen die de onzekerheid van een meting en een voorspelling aangeven. De grafieken met weersvoorspellingen zijn daar een mooi voorbeeld van. Metten is dus een tamelijk bescheiden onderneming. De meeste van onze persoonlijke meningen en conclusies zijn heel wat minder bescheiden en ze gaan zelden vergezeld van betrouwbaarheidsintervallen. Cijfertjes zijn natuurlijk niet alles. Zo moeten

we ervoor uitkijken dat we het onderwijs niet zozeer gaan overladen met allerlei toetsen dat scholen hun beleid niet volledig laten afhangen van toetsscores dat er te weinig tijd over blijft voor het primaire proces, het leren. Net als voor de meeste dingen op deze aarde geldt ook hier: alles met mate.

Cees Glas studeerde Psychometrie en Statistiek aan de Universiteit Utrecht. Van 1982 tot 1995 werkte hij als methodoloog en statisticus bij het CITO. Hij is als hoogleraar verbonden aan de Universiteit Twente en hoofd van de vakgroep Onderzoeksmethoden, Meetmethoden en Data-analyse. Hij treedt vaak op als adviseur bij grootschalig onderwijskundig onderzoek.

Verhoeven Meten is zweten!

Een paar uur voordat ik aan deze column begon, maakte de Kiesraad de officiële uitslag van de verkiezingen van 12 september 2012 bekend. Hieruit valt een schat aan informatie te halen. Zo haalde D66 757.091 stemmen, ruim 8% van het totaal uitgebrachte aantal; ruim honderdduizend meer dan de 654.167 stuks (6,7%) van twee jaar geleden! Dit levert ons twee extra zetels op in de Tweede Kamer. Naast lijsttrekker Alexander Pechtold en eerste vrouw op de lijst Stientje van Veldhoven werd ook Pia Dijkstra op voorkeurstemmen gekozen. Zij haalde 24.886 stemmen. Ga er maar aan staan.

Ja, metten is weten. En er is meer gemeten dat u moet weten.

Uit onderzoek van opiniepeiler Maurice de Hond en geograaf Josse de Voogd blijkt dat D66 in de grote (universiteits)steden de hoogste stempercentages haalde, namelijk tussen de 10 en 15%. Maar ook in Brabant en Gelderland hebben we het relatief goed gedaan. Sowieso doen we het buiten de Randstad steeds beter. Minder positief is dat wij het slecht doen onder vrouwen. De verhouding man-vrouw is ongeveer 60-40. Er is geen partij met een schevere verhouding. Verder blijkt dat mensen vooral op ons stemmen vanwege onze standpunten over Europa, onderwijs en – in mindere mate – de overheidsfinanciën.

Metten is weten, ja. Maar de volgende weetjes zijn pas echt veelzeggend.

Namelijk dat we onze kiezers verrassend goed weten vast te houden: bijna tweederde van de D66-stemmers uit 2010 bleef ons trouw. Daarbij heeft liefst 20% van onze kiezers pas op de verkie-

zingsdag besloten D66 te stemmen. Geen andere partij had zoveel last-minute stemmers. En 27% van de D66-stemmers heeft uit strategische overwegingen op onze partij gestemd.

Weet u, die laatste cijfers stemmen positief. We zijn er in onze campagne blijkbaar in geslaagd tot op het laatste moment mensen te overtuigen. Ondanks de losgebarsten en opgeklopte premierstrijd tussen Samsom en Rutte. Ook al was het niet makkelijk, we zijn tot het bittere einde doorgedaan. Met zijn allen. In het hele land. Eensgezind en strijdlustig. Dat is de grootste winst die we konden boeken. En daar wil ik iedereen voor danken: Alexander en alle andere kandidaten, het bestuur, het campagneteam, alle medewerkers, de afdelingen en bovenal de vrijwilligers. Goed werk!

Toch zijn er ook minder positieve geluiden. Want zonder die tweestrijd tussen PvdA en VVD had D66 zelfs 14 zetels gehaald. En in de peilingen hebben we zelfs een paar keer boven de 15 gestaan...

De peilingen. Een Haags gezegde luidt: peilingen zijn palingen. Een peiling is immers een steekproef met een statistische foutmarge. Geeft hooguit een trend weer. Nou, u mag best weten dat een daling in de peilingen onrustige uurtjes oplevert voor de campaigners. Tegen beter weten in. Want als je daalt, is dat nieuws dat zichzelf versterkt. En voor je het weet zit je in een negatieve spiraal. Die nergens op gebaseerd is, maar waar iedereen nerveus van wordt.

En dan is metten opeens zweten. Daarom trek ik de volgende conclusie:

‘Metten is metten.
Weet wat je moet weten
En wat je snel weer moet vergeten.’

Kees Verhoeven, Tweede Kamerlid voor D66 en campagneleider tijdens de afgelopen Tweede Kamer-verkiezingen