
Terugval naar het gemiddelde; een methodologische notitie

Jos Dessens en Wim Jansen*

In het navolgende wordt een valstrik besproken, van welks bestaan beoefenaren van de sociale wetenschappen al enkele decennia op de hoogte zijn of kunnen zijn. In de praktijk van het empirisch onderzoek blijkt soms het tegendeel: causale processen worden gepostuleerd daar, waar slechts sprake is van artefacten. In deze korte bijdrage willen we enkele voorbeelden daarvan onder de loop nemen. Dit doet af aan de kwaliteit van het betreffende onderzoek. Wanneer artefacten worden gebruikt als ondersteuning van beleid, kunnen de gevolgen nog vervelender zijn. Bovendien draagt dit niet bij tot een positieve beeldvorming van de sociale wetenschappen bij het publiek. We vinden het daarom van belang om bij wijze van illustratie én waarschuwing, nog eens aandacht te besteden aan één der valkuilen in de moeizame weg van de sociaal-wetenschappelijke onderzoeker.

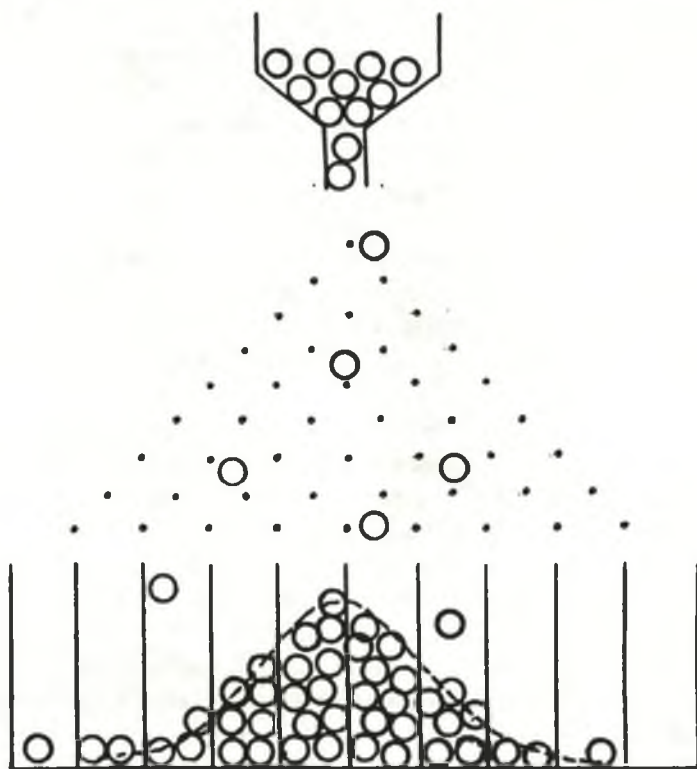
Een van de grootste moeilijkheden bij een vliegopleiding is het leren maken van een zachte landing. Al weer enige jaren geleden meenden instructeurs, die verbonden waren aan een opleidingscentrum voor piloten, een belangrijke pedagogische wetmatigheid te hebben ontdekt. Zij merkten op, dat het uitvoerig prijzen van leerlingen, die een perfecte landing uitvoerden, in het algemeen leidde tot een slechtere landing op een later tijdstip, terwijl het uitfoeteren van leerlingen die een slordige landing uitvoerden, in het algemeen leidde tot een betere prestatie op een later tijdstip. De conclusie ligt voor de hand: 'Prijs mensen nooit om hun goede prestaties, maar bestraf hen bij slechte prestaties' (Kahneman & Tversky, 1973; Tversky & Kahneman 1974).

Men zou zich echter kunnen afvragen, of het achterwege laten van zowel uitvoerige loftuitingen als berispingen tot een ander resultaat zou leiden, met andere woorden, of het belonende c.q. bestraffende gedrag van de instructeurs in verband staat met de prestaties van de leerlingen. Ter beantwoording beginnen we met een model te maken voor variaties in de prestaties van één individu. Prestaties zijn het resultaat van, natuurlijk, het algemene prestatieniveau, maar daarnaast ook van een groot aantal, min of meer onafhankelijke, min of meereven

* Met dank aan W. Kremers, P. G. Swanborn en A. Verbeek voor hun commentaar.

grote effecten. Zo wordt de prestatie van een hardloper in eerste instantie bepaald door aanleg, training en conditie, maar daarnaast door allerlei factoren als windsnelheid, buitentemperatuur, mate van concentratie, staat van zijn uitrusting, aanwezigheid van toeschouwers, etc. Gegeven dat deze factoren qua effect van dezelfde orde van grootte, additief en onafhankelijk van elkaar zijn, zullen herhaalde prestaties een normale verdeling (vgl. Centrale Limiet Stelling) volgen met als gemiddelde het prestatieniveau.

We kunnen dit illustreren aan de hand van het bord van Galton (zie figuur 1).



Figuur 1. Bord van Galton

De balletjes vallen op de pinnetjes; de kans om links of rechts van een pinnetje af te springen is gelijk. Dit naar links of rechts afspringen zouden we kunnen vergelijken met het wel of niet positief uitvallen van een bepaald effect van een bepaalde factor op de prestatie. Voor de balletjes, die in het uiterst linkse of rechtse vakje terecht zijn gekomen, geldt, dat ze steeds links c.q. rechts van een pinnetje zijn afgesprongen. Vertaald naar de door onze atleet te leveren prestatie: alle factoren

zaten 'toevallig' tegen of alle factoren zaten 'toevallig' mee.

Ter completering van het antwoord voeren we nu het volgende gedachtenexperiment uit: we merken de balletjes, die in de extreme vakjes terecht zijn gekomen, met een kleur; vervolgens laten we alle balletjes opnieuw op de pinnetjes vallen. De gekleurde balletjes zullen niet opnieuw in de extreme vakjes terecht komen, maar zich daarentegen willekeurig over alle vakjes verdelen. Op grond van het toeval zit het er niet in, dat dezelfde balletjes steeds in dezelfde (extreme) vakjes terecht komen. Wel zal de proportie balletjes, die in de extreme vakjes terecht komt van keer tot keer ongeveer even groot zijn: *de (prestatie)verdeling verandert niet*. Dit laatste is belangrijk genoeg om op te merken, omdat men wellicht zou vermoeden, dat de prestatieverdeling een steeds geringere spreiding rond het gemiddelde zou gaan vertonen. Dit is (zoals het experiment met het bord van Galton aantoont) niet zo: minder extreme gebeurtenissen kunnen gevolgd worden door extremere gebeurtenissen, waardoor de uiteindelijke vorm van de verdeling niet verandert. Men zou wellicht tegen kunnen werpen, dat het telkens weer laten vallen van balletjes op pinnetjes onafhankelijke herhalingen zijn, terwijl dit voor de prestaties van piloten niet het geval is. We maken daarom een onderscheid tussen het 'ware' prestatieniveau en de afwijkingen hierop ten gevolge van toevallige verstoringen. De 'ware' prestaties zijn bij herhaling uiteraard zeer afhankelijk; de toevallige verstoringen zijn daarentegen vrijwel onafhankelijk.

We hebben nu tevens het mechanisme bloot gelegd, dat ten grondslag ligt aan de pedagogische misvatting waar we mee zijn begonnen: extreme gebeurtenissen (harde c.q. zachte landingen) zijn gebeurtenissen, waarbij pech dan wel geluk een belangrijke rol speelt. Bij herhaling zal men in het algemeen niet opnieuw dezelfde mate van geluk of pech hebben. Uitzonderingen vinden we in fictie: Guus Geluk in de Donald Duck-strips. In de literatuur staat dit verschijnsel bekend als *regressie (terugval) naar het gemiddelde*.

Wat voor prestaties van individuen geldt, geldt ook voor prestaties van groepen van individuen. De variabele 'groepsprestatie' is samengesteld uit de individuele, normaal verdeelde prestatiescores en zal dus eveneens normaal verdeeld zijn. Ook dan kan er sprake zijn van terugval naar het gemiddelde. Een zwemvereniging, die de vier beste zwemsters uit de interne selectiewedstrijden zonder extra training afvaardigt naar de regionale kampioenschappen, zal merken dat deze zwemsters daar op de estafette vaak een slechtere prestatie (totaaltijd) leveren. Ook in dit geval wijst men dan veelal factoren als 'nervositeit', 'onbekend terrein' e.d. als oorzaken aan, maar zeker is dat terugval naar het gemiddelde hier een voorname, veelal onderschatte rol speelt.

Het herkennen van situaties waarin terugval naar het gemiddelde een rol heeft gespeeld blijkt vaak moeilijk. Het kenmerkende van die situaties is, dat geselecteerd wordt op een variabele, die van meting tot meting per waarnemingseenheid

behoorlijk kan variëren. Door het selecteren wordt het effect van terugval naar het gemiddelde geïntroduceerd. Uit de literatuur blijkt, dat zelfs ervaren onderzoekers de interpretatie van terugval naar het gemiddelde veronachtzamen. In de VS onderzochten Rosenthal & Jacobson (1968) de invloed van de verwachting van de onderwijzers met betrekking tot de prestaties van de leerlingen op die prestaties. Zij deden dit door op basis van het toeval namen van zogenaamde 'spurters' of 'bloomers' aan het onderwijzend personeel mede te delen en na enige tijd de toename in IQ (= posttest IQ minus pretest IQ) bij de groep 'spurters' en de groep 'non-spurters' vast te stellen. Inderdaad bleek de toename in IQ voor de eerstgenoemde groep kinderen veel groter dan voor de laatstgenoemde groep, vooral in de eerste klasse (Rosenthal & Jacobson, tabel 7.1, p. 75). Uit tabel A-1 (p. 187) blijkt echter, dat het gemiddelde IQ op de pre-test van de als 'spurters' aangemerkte eerste-klassers in twee van de drie klassen behoorlijk lager was dan het gemiddelde IQ op de pretest van de overige kinderen in de respectieve klassen. Als we er van uit gaan, dat dit verschil toevallig is, dan zullen de als 'spurters' aangemerkte kinderen al een bij voorbaat grotere toename in de IQ-score te zien geven. Deze interpretatie in termen van terugval naar het gemiddelde ondergraaft de conclusies van Rosenthal & Jacobson, al weten we niet in hoeverre.

In zijn studie naar oorzaken van machtsaspiraties toetste Veen (1970, p. 180 e.v.) een tweetal hypothesen over de relatie tussen 'grootte van de machtsafstand' en 'prestatie'. Een van deze hypothesen luidde: 'Naarmate de machtsafstand kleiner wordt zal de prestatie verbeteren'. De variabele 'grootte van de machtsafstand' werd in dit veldexperiment gemanipuleerd via toewijzing van de proefpersonen aan een participatieconditie 'P' (het trainingsprogramma werd opgesteld door de trainer in overleg met de groep) of 'NP' (het trainingsprogramma werd opgesteld zonder overleg met de groep; het zelfde programma als in de participatieconditie is opgesteld). De variabele 'prestatie' werd geoperationaliseerd als de snelheid waarmee een hockey-hindernisbaan wordt afgelegd, waarbij fouten strafseconden opleveren. De prestatie werd op vier tijdstippen gemeten: aan het begin van het experiment, voorafgaand aan de training (t_0) en na resp. 4, 6 en 52 weken (t_4 , t_6 en t_{52}). Er werd gematched op de variabelen 'leeftijd', 'Individuele P (rominentie)-score' en 'hockeycapaciteiten'. Matching op deze laatste variabele geschiedde echter niet op basis van de prestatie geleverd op t_0 , maar op basis van een beoordeling door de onderzoeker zelf (id. p. 146). Van random toewijzing van de proefpersonen aan de verschillende condities was geen sprake. De resultaten (in seconden) van prestatie naar wel/niet participatie kunnen als volgt worden weergegeven:

	conditie	
	P	NP
t_0	25.2	23.5
t_4	20.2	21.9
t_6	20.1	24.5
t_{52}	17.9	18.5

Bij de toetsing van zijn hypothesen lette Veen vooral op de verschillen in prestatie tussen t_6 en t_0 , hetgeen Veen tot bevestiging van zijn hypothese doet besluiten (zie tabel). Welnu, waarschijnlijk kunnen de gevonden verschillen uit het effect van 'terugval naar het gemiddelde' worden verklaard; al dan niet meebeslissen over de opzet van de training heeft geen effect op de prestatie. Een extra aanwijzing voor de juistheid van deze veronderstelling is het verwaarloosbaar kleine verschil in prestatie tussen de beide condities op t_{52} ; dit kan beschouwd worden als het gestabiliseerde effect van een jaar training.

Ook een aantal resultaten van een evaluatieonderzoek naar de effecten van een bedrijfsopleiding (Van der Vegt, 1974) kunnen geheel in termen van 'terugval naar het gemiddelde' worden verklaard. In dit onderzoek werden aan managers, verdeeld over een drietal condities (school A, school B, school C) op een drietal tijdstippen (aan het begin van de school: t_0 ; aan het eind van de school, d.w.z. na 4 weken: t_1 ; 6 à 7 maanden na afloop van de school: t_3) een aantal opinie- en houdingsvariabelen gemeten. Van randomisering van de proefpersonen over de verschillende condities was geen sprake, wel heeft Van der Vegt zich veel moeite getroost om na te gaan of de groepen op een groot aantal mogelijk verstorende variabelen significant verschilden (id. p. 114 e.v.). Ons commentaar betreft de resultaten, zoals weergegeven in de tabellen II.1a (id. p. 141), II.1b (id. p. 145), II.4 (id. p. 146), II.5 (id. p. 160) en II.7 (id. p. 165). Voor de variabelen 'opvattingen over leidinggeven', 'toegekende potentie Human-Relations-training' en 'toegekend belang Human-Relations-aspecten' (vergelijk de drie laatst genoemde tabellen) geldt voor alle groepen een hoge score bij de eerste meting. We zien dan ook op een tweede en/of derde meetmoment een terugval naar het – lager gelegen – 'ware' gemiddelde. Met betrekking tot de variabelen 'beeldvorming over functies per factor' en 'beeldvorming over functies totaal' (vergelijk resp. de tabellen II.1b op p. 145 en p. 146) geldt een waarschijnlijk lager-dan-het-'ware'-gemiddelde score op t_1 , hetgeen *dus* leidt tot een toename op t_2 . De variabele 'mate van stereotyperende beeldvorming over managers' (vergelijk tabel II.1a) werd al in afwijkingenscores gemeten. We zien hier een bevestiging van de stelling: hoe groter de absolute afwijking van het gemiddelde, hoe groter het terugval-effect.

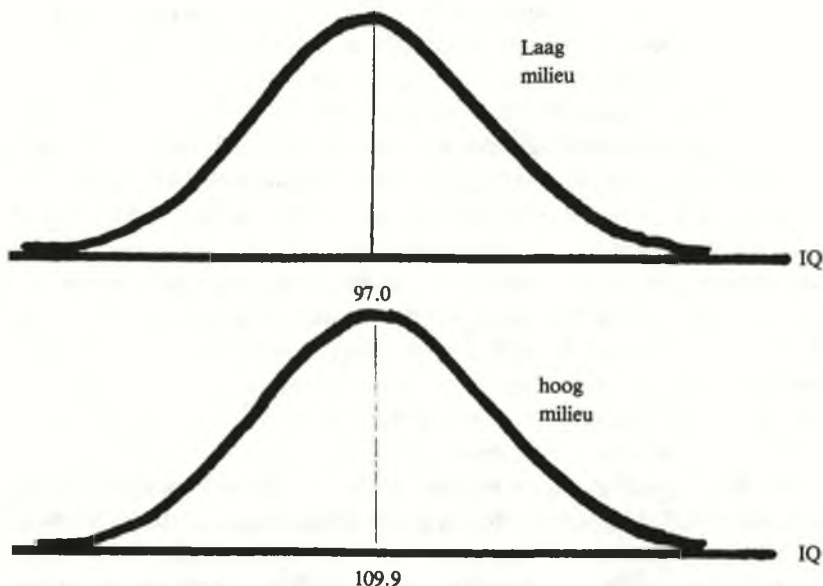
Peschar deed in 1975 een onderzoek naar de invloed van het sociaal milieu op het uiteindelijk bereikte opleidings- en beroepsniveau. Hij had de beschikking over data met betrekking tot onder andere de variabelen sociaal milieu, leeftijd, sekse en IQ van kinderen uit de stad Groningen ten tijde van de afname van een schoolgeschiktheidstest (GALO) in 1958/59 en 1959/60. Op basis van dit materiaal selecteerde Peschar steeds paren leerlingen, waarbij de een uit een hoog milieu en de ander uit een laag milieu afkomstig was, terwijl binnen het paar sekse, leeftijd en IQ gelijk waren. Na 15 jaar spoorde Peschar de leerlingen weer op en vroeg hen onder andere naar het uiteindelijk bereikte opleidings- en beroepsniveau. Het gemiddeld opleidings/beroepsniveau van de groep leerlingen, afkomstig uit het hoge sociaal milieu werd vergeleken met dat van de groep leerlingen, afkomstig uit het lage sociaal milieu. Omdat de paren steeds gelijk waren ('gematched') op sekse, leeftijd en IQ, konden eventuele verschillen in de gemiddelden niet geweten worden aan verschillen in deze variabelen, maar aan het verschil in sociaal milieu. Peschar vond inderdaad verschillen, en concludeerde dat het sociaal milieu de verantwoordelijke factor is.

In een reactie op de publikatie van de onderzoeksresultaten stellen De Groot en Van Peet (1975), dat Peschar in de kuil 'terugval naar het gemiddelde' is gevallen. Zien we nog even af van de formules, die zij gebruiken om het effect van de terugval te berekenen, en proberen we hier duidelijk te maken waarom hier een terugval naar het gemiddelde een rol gespeeld kan hebben.

Groepen leerlingen uit laag en hoog sociaal milieu zullen een verschillend gemiddelde IQ hebben. Dat is wat Peschar ook vindt: hoog milieu: gemiddeld IQ: 109.9, $Sd=14.9$; laag milieu: gemiddeld IQ: 97.9, $Sd=13.4$. Bij een grafische weergave van het bovenstaande (zie figuur 2) valt op, dat ten gevolge van de matchingsprocedure bij de te vormen paren ('matches') degenen uit het lage sociaal milieu vooral *boven* het gemiddelde IQ van hun groep zitten, terwijl de leerlingen uit het hoog sociaal milieu vooral *onder* het gemiddelde IQ van hun groep zitten. Voor IQ-scores geldt hetzelfde als voor prestatiescores. Voor het hoger milieu geldt, dat nogal wat leerlingen een te lage (dat wil zeggen lager dan hun 'ware') IQ-score zullen hebben; voor het lage milieu geldt, dat nogal wat leerlingen een te hoge (dat wil zeggen hoger dan hun 'ware') IQ-score zullen hebben: binnen een 'gematched' paar is wel degelijk een verschil in 'ware' IQ aanwezig. Terugval naar het gemiddelde heeft hier onmiskenbaar een rol gespeeld.

Volledigheidshalve vermelden wij in deze context, dat Peschar met argumenten aannemelijk maakt, dat terugval naar het gemiddelde in zijn onderzoek geen belangrijke rol kan hebben gespeeld. Allereerst toont hij via berekening van het terugvaleffect overtuigend aan, dat zijn conclusies maar in zeer geringe mate worden aangetast. Daarnaast wijst hij op het feit, dat het oordeel van de onderwijzer bij een aan te geven advies over voortgezet onderwijs gebaseerd is op de gemeten IQ-score en niet op de 'ware'

IQ-score. De gemeten IQ-score fungeert als etiket, dat mogelijk als een zichzelf waarmakende voorspelling fungeert'.



Figuur 2. Verdeling van IQ voor laag resp. hoog milieu

Ook bij beleidsvoering loopt men het risico in de valkuil te stappen. Nietongebruikelijk is het immers om pas in te grijpen, wanneer een bepaalde grens wordt overschreden. Bij voorbeeld wanneer in een bepaald jaar het aantal verkeersslachtoffers een sterke toename te zien geeft. Indien nu de overheid besluit hier iets aan te doen door bij voorbeeld een maximumsnelheid op autosnelwegen in te voeren, dan is het verleidelijk om de in het volgend jaar welhaast gegarandeerd optredende daling van het aantal verkeersslachtoffers geheel toe te schrijven aan deze maatregel, terwijl een verklaring in termen van terugval naar het gemiddelde plausibeler is.

In de zestiger jaren is er gediscussieerd over de vraag naar de effectiviteit van compensatieprogramma's. Kinderen uit kansarme milieu's met een naar verhouding lage score op bij voorbeeld taalvaardigheid krijgen extra begeleiding. Bij de evaluatie van het programma is niet meer uit te maken, of een eventuele toename in taalvaardigheid nu te wijten is aan dat programma, dan wel aan terugval naar het gemiddelde. Waarschijnlijk aan beide. Om het 'netto'-effect van het compensatieprogramma vast te stellen zouden we kinderen uit kansarme milieu's willekeurig toe moeten wijzen aan een groep, die extra begeleiding krijgt, of een groep, die dat niet krijgt. Weinig onderzoekers zullen met een dergelijke procedure akkoord gaan.

Een procedure waardoor in dit kader een ideaal experiment beter wordt benaderd, is het selecteren van kinderen, die in aanmerking komen voor compensatieprogramma's op basis van het feit dat de school waarop ze zitten zich in een 'achterstandsbuurt' bevindt. Bovendien wordt een controlegroep samengesteld van kinderen uit dezelfde buurt, maar die *niet* in deze buurt op school gaan. Deze opzet werd in de Verenigde Staten bij de zogenaamde Head Start programma's toegepast (Campbell & Erlebacher, 1975). De resultaten kunnen dan heel opmerkelijk zijn: de experimentele groep blijft gelijk (het programma heeft geen effect) of geeft een toename in de prestatie te zien (het programma heeft effect); de controlegroep daarentegen vertoont een *daling* met betrekking tot de prestatie. Naïeve of kwaadwillige beoordelaars kunnen hieraan onjuiste conclusies met betrekking tot het beleid verbinden. De grondslag voor een dergelijke misvatting ligt in het feit, dat de experimentele groep *niet* is geselecteerd op basis van *individuele* (extreem) lage prestatiescores, terwijl de controle-groep een, door de onderzoeker geselecteerd, hoger niveau kan hebben. Dit is des te meer waarschijnlijk naarmate de selectievariabelen sterker gecorreleerd zijn met de 'school'-variabelen.

Het overeenkomstige in de voorgaande voorbeelden is, dat de onderzoekers niet in staat waren de onderzoeksgroepen at random samen te stellen. We dienen dan vooral bedacht te zijn op de vraag of de experimentele en de controlegroep niet reeds *bij voorbaat* verschillen op mogelijk verstorende variabelen. In dat geval ligt een alternatieve interpretatie in termen van terugval naar het gemiddelde voor de hand.

De onderzoeker echter die meent de onmogelijkheid van randomisering te ondervangen door te matchen op een aantal mogelijk verstorende variabelen valt daarmee gegarandeerd in de valkuil van terugval naar het gemiddelde: matchen zonder randomisering is uit den boze!

In het dagelijks leven kunnen we nog een aantal situaties aantreffen, waar voor gebeurtenissen allerlei populaire verklaringen worden gegeven, maar verklaring in termen van terugval naar het gemiddelde meer voor de hand ligt. Zo meldt de 9-jarige Suzanne, dat ze alsmear probeert haar record kaatsenballen (56 keer) te verbeteren, en dat dat maar niet wil lukken (er is een 'wachttijd' tot een nieuw maximum). De briljante middelbare scholier krijgt van zijn ouders op zijn kop, omdat hij voor de laatste repetitie natuurkunde géén 10 haalde zoals bij de vóórlaatste keer ('je hebt er zeker met je pet naar gegooid, hè?'¹). Ook hebben mensen de behoefte om, weliswaar onwaarschijnlijke, maar zeer aangename gebeurtenissen opnieuw te beleven; uiteraard volgen reacties van teleurstelling ('het is nooit meer zoals die éne keer'). Het omgekeerde komt ook voor: het vermijden van als zeer onaangenaam – hoewel zeer onwaarschijnlijke – beleefde situaties als trein- en vliegtuigongelukken door niet meer per trein of vliegtuig te reizen (een navrante illustratie van de naïeve gedachte, dat het gekleurde balletje

telkens weer in het extreme vakje terecht komt). Onvoldoende rapportcijfers kunnen ook zonder bijlessen verbeteren. Hoge bloeddruk kan ook zonder pillen bij een volgende meting een normalere blijken.

Het probleem van terugval naar het gemiddelde zou de wereld uit zijn, indien we zouden beschikken over de 'ware' scores, dat wil zeggen dat we in staat zijn zonder meetfouten te meten. Dit nu is vaak onmogelijk. Wel kunnen we er over denken om deze 'ware' scores te schatten op basis van de waargenomen scores. Dit schatten van de 'ware' scores op basis van de waargenomen scores is alleen mogelijk indien we beschikken over kennis omtrent de betrouwbaarheid van het meetinstrument. De betrouwbaarheid kan bij voorbeeld vastgesteld worden door middel van herhaald meten (test-hertest betrouwbaarheid), dan wel door opsplitsing van het meetinstrument (halveringsbetrouwbaarheid). Een andere mogelijkheid is, dat de betrouwbaarheid van het meetinstrument in eerder onderzoek reeds is vastgesteld. Dit is bij voorbeeld het geval bij de Stanford-Binet intelligentietest en de GALO-test. Met behulp van de volgende notatie kunnen we de schattingsmethode beschrijven. De 'ware' score op een variabele X noemen we $X_{t(rue)}$, de waargenomen score $X_{o(bserve)}$ en het verschil tussen 'ware' en waargenomen $X_{e(error)}$: $X_o = X_t + X_e$. Afgeleid kan worden, dat de 'ware' score met behulp van lineaire regressie als volgt kan worden geschat (Lord and Novick, 1974, p. 65):

$$\hat{X}_t = r_{XX1} X_o + (1 - r_{XX1})X_o$$

In deze formule is r_{XX1} de betrouwbaarheidscoëfficiënt van het meetinstrument in kwestie en X_o de gemiddelde groepsscore. Stel dat de betrouwbaarheid van het meetinstrument $r_{XX1} = .80$ dan is bij een score $X_o = 100$ een goede schatting van de 'ware' score:

$$\hat{X}_t = (.80)(100) + (1 - .80)(75) = 95$$

Toegepast op de situatie bij Peschar: voor een 'gematched' paar met een gemeten IQ-score van 130 kunnen we, uitgaande van een betrouwbaarheid van .80 voor de GALO-test, voor de persoon uit het lage milieu een 'ware' IQ-score voorspellen van:

$$(.80)(130) + (1 - .80)(97) = 123.4$$

en voor een persoon uit het hoge milieu:

$$(.80)(130) + (1 - .80)(109.9) = 126.0$$

Een verwacht verschil in IQ van 2.6 punten tegenover een gemeten verschil van 0 punten.

Beschikken we dus over een schatting voor de betrouwbaarheid van het gehanteerde meetinstrument, dan zijn we in principe in staat het effect van het verschijn-

sel terugval naar het gemiddelde te schatten, door gebruik te maken van de voorspelde 'ware' scores in plaats van de waargenomen scores.²

Noten

1. Hoewel beide voorgaande voorbeelden verwijzen naar prestatie-meting, is er een essentieel verschil: bij rapportcijfers kan het terugvaleffect mede veroorzaakt zijn door het plafond in de meetschaal.
2. Zo zou Peschar bij het samenstellen van de paren leerlingen beter gebruik hebben kunnen maken van de voorspelde 'ware' IQ-scores in plaats van de waargenomen IQ-scores. In een reactie op De Groot en Van Peet voert Peschar deze correctie uit, waarbij hij nagaat wat de consequenties hiervan zijn voor zijn eerdere conclusies.

Geraadpleegde literatuur

- Campbell, D. T. & A. Erlebacher, How regression artifacts in quasi-experimental evaluations can mistakenly make compensatory education look harmful, in: E. L. Stroening & H. Gutentag (eds.), *Handbook of evaluation Research*, Vol. I, 1975, Beverly Hills, ISBN 0-8039-0428-2.
- Groen, H. K., Leerlingen uit verschillende sociale milieu's hebben geen gelijke kansen, *Tijdschrift voor Onderwijsresearch*, 1, 1975, pp. 40-42.
- Groot, A. D. de & A. A. J. van Peet, Enkele kanttekeningen bij het proefschrift van J. L. Peschar: milieu, school en beroep, *Tijdschrift voor Onderwijsresearch*, 1, 1975, pp. 36-39.
- Groot, A. D. de & A. A. J. van Peet, Nogmaals: de invloed van regressie-effecten, *Tijdschrift voor Onderwijsresearch*, 1, 1975, pp. 133-137.
- Kahneman, D. & A. Tversky, On the psychology of prediction, *Psychological Review*, 80, 1973, pp. 237-251.
- Lord, F. M. & R. M. Novick, *Statistical theories of mental test scores*, 1968, 1974 (2nd), Reading, ISBN 0-2010-4310-6.
- Peschar, J. L., *Milieu, school, beroep*, 1975, 1978 (2e aangevulde en verbeterde druk), Groningen, ISBN 9-0629-3001-8.
- Peschar, J. L., De invloed van regressie-effecten in het milieu-school-beroep-onderzoek: een antwoord aan A. D. de Groot en A. A. J. van Peet, *Tijdschrift voor Onderwijsresearch*, 1, 1976, pp. 49-58.
- Peschar, J. L., Andermaal de invloed van regressie-effecten, *Tijdschrift voor Onderwijsresearch*, 1, 1976, pp. 137-138.
- Rosenthal, R. & L. Jacobson, *Pygmalion in the classroom. Teachers expectation and pupil's intellectual development*, 1968, Holt, Rinehart and Winston Inc., New York.
- Tversky, A. & D. Kahneman, Judgement under uncertainty: heuristics and biases, *Science*, 185, 1974, pp. 1124-1131.
- Veen, P., *Meebeslissen. Een veldexperiment in een hockeyclub*. diss., 1970, Utrecht.
- Vegt, R. van der, *Opleiden en evalueren. Een veldexperimentele studie naar uitkomsten van een bedrijfsopleiding*, 1974, Meppel, ISBN 9-0600-9120-5.