# Over de kwaliteit van het bepalen van leerlingprestaties: de 4 D's

Ook beoordelen voor het PTA is een vorm van evalueren van onderwijs. Evalueren bestaat uit drie stappen. Vaststellen wat leerlingen feitelijk gedaan hebben (meten), een waardering uitspreken over wat een leerling gedaan heeft (waarderen, en de ene prestatie is de andere niet) en communiceren over die waardering van een bepaalde 'prestatie'. In een simpel schema:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **meten** | **→** | **waarderen** | **→** | **rapporteren** |

Belangrijk is dat er een principieel onderscheid is tussen het meten van een prestatie en het waarderen daarvan en dat het rapporteren over die waardering ook weer echt een aparte fase is. Eigenlijk maakt de rapportage niet eens deel uit van de evaluatie (maar wel van het hele proces van evalueren), maar niet zelden worden deze drie afzonderlijke fasen met elkaar verward, door elkaar gehanteerd of anderszins onvoldoende gescheiden.

Om te kunnen evalueren moet je dus informatie verzamelen over wat leerlingen precies gedaan hebben, iets ‘meten’zou je dat kunnen noemen. Elke vorm van evalueren maakt op een of andere manier gebruik van meetmethoden. En aan dat meten moet je bepaalde eisen stellen om te waarborgen dat je evaluatie waardevol is, of andersom, om te voorkomen dat die zonde van de moeite is. De mate waarin aan die eisen wordt voldaan hangt af van de gebruikte methode en niet zozeer van het te evalueren kwaliteitsaspect van bewegingsonderwijs. De eisen zijn: validiteit, betrouwbaarheid, hanteerbaarheid en transparantie. Als ezelsbruggetje duiden we die eisen ook wel aan als de 4 D’s: *dekkend*, *docent-onafhankelijk*, te *doen* en *doorzichtig* (zie figuur 1).



*Figuur 1 De 4 D’s: eisen aan goed meten*

**Dl: Dekkend (validiteit)**

De validiteit van een meetmethode wordt bepaald door de mate waarin een methode werkelijk meet wat hij zou moeten meten. Hier gaat het in wezen om de goede beantwoording van de wat-vraag. Zegt wat je meet wel iets over het *'enig aspect van bewegingsonderwijs'* dat je wilde evalueren? Men zegt ook wel: dekt datgene dat gemeten wordt de doelstelling wel? Vandaar dekkend.

Soms ligt dat tamelijk eenvoudig. Als je wilt weten hoe snel iemand kan lopen, laat je die persoon een stukje zo hard mogelijk lopen, je meet de tijd die hij daarover doet en dan is dat een valide meting. Meestal ligt het echter niet zo simpel. Er kunnen twee soorten problemen zijn. Stel dat je wilt weten hoe goed iemand kan badmintonnen. ledereen zal het erover eens zijn dat een factor als spelinzicht of een aspect als tactisch slim lang-kort spelen dan van belang is om te meten. Maar hoe kun je dat nu ‘meten’? Als je daarvoor probeert een dekkende meting te formuleren kom je – als je echt kritisch bent – vaak hopeloos in de problemen. De Groot (1986) noemt dat het dekkingsprobleem. Weet je zeker dat je op een of andere manier alleen spelinzicht meet en niet iets anders tegelijkertijd? De vraag is dan of het gebruikte meetinstrument, bijvoorbeeld een observatielijst, inderdaad, en alleen maar, spelinzicht meet.

Er zijn doelstellingen die zich niet goed vooraf laten operationaliseren, bijvoorbeeld als over doelstellingen wordt gesproken in termen van competenties. Soms helpt het om te proberen je doelen te vertalen in heel concrete evaluatiedoelstellingen omdat je daardoor gedwongen wordt om preciezer te zeggen wat je nu eigenlijk beoogde. Soms is het zelfs de vraag of je dat wel moet proberen, vooral bij sterk persoonsgebonden doelen of bij doelen die gaan over hoe leerlingen bepaalde activiteiten beleven. Vanuit een handelingstheoretisch vakconcept is het extra lastig, omdat je eigenlijk wilt redeneren vanuit het handelingsperspectief van elke leerling afzonderlijk en dan eigenlijk ook per leerling individuele doelen zou moeten omschrijven.

Een ander soort validiteitsprobleem ontstaat als je wel goed kunt meten wat je wilt meten, maar niet meer weet of hetgeen je meet wel van belang is voor de evaluatievraag die je wilde beantwoorden. Daarvoor gaan we even terug naar het voorbeeld van badminton. Als je voor de evaluatie voor het bijstellen van een lessenreeks badminton bijvoorbeeld alleen zou meten hoe hard de leerlingen aan het eind van die reeks een ‘badminton-pluimpje kunnen slaan’, dan voelt iedereen op zijn klompen aan dat je daarmee nooit een antwoord krijgt op de vraag wat je moet bijstellen aan die lessenreeks. Daar speelt vast veel meer een rol. Nu ligt het meestal niet zó klip en klaar, maar de aard van het probleem is duidelijk. Helaas zijn het vaak de relevante aspecten die zich lastig valide laten evalueren, terwijl de makkelijk en tamelijk objectief te meten aspecten vaak minder relevant zijn ten opzichte van de doelstelling.

**D2: Docent-onafhankelijk (betrouwbaarheid)**

Behalve valide moet een meetmethode ook betrouwbaar zijn. Betrouwbaar wil zeggen dat een meetmethode eenzelfde ‘iets’ (eigenschap, kwaliteit) telkens met hetzelfde resultaat meet, als dat iets niet verandert. Twee vormen van betrouwbaarheid van een meting zijn hier van belang: herhalingsbetrouwbaarheid en interbeoordelaarsbetrouwbaarheid.

Herhalingsbetrouwbaarheid, doorgaans test-hertest-betrouwbaarheid genoemd, is de mate waarin scores die op twee tijdstippen wordt gemeten overeenstemmen. Herhalingsbetrouwbaarheid gaat om de vraag of het verschil uitmaakt wanneer je de meting doet. Scoort iemand vandaag hetzelfde als volgende week? Want als dat niet zo is moet je ernstig uitkijken wannéér je gaat beoordelen. Storingen kunnen zitten in de beoordelingssituatie (stel dat je net tegen de beste speler van de klas moet badmintonnen) of in de momentele kenmerken van de beoordelaar (de docent heeft net lesgegeven aan een vervelende klas en is flink chagrijnig) of de beoordeelde (een leerling is erg moe na een feestje of juist in grote vorm omdat hij verliefd is). Uiteraard is het de grote truc om er achter te komen wat de ‘juiste score’ is die je wilt hebben. De enige oplossing is een paar keer te meten en het gemiddelde te nemen van de scores of om in elk geval beoordelingen waarvan je vermoedt dat ze onder 'verdachte omstandigheden' zijn gedaan niet mee te tellen.

Interbeoordelaarsbetrouwbaarheid is de mate waarin een meetprocedure bij twee beoordelaars steeds hetzelfde resultaat oplevert. Interbeoordelaarsbetrouwbaarheid wil zeggen of het wat uitmaakt welke docent staat te beoordelen, een ‘rekkelijke’ of een ‘precieze’. Misschien ken je dat (nog) wel, dat twee docenten aan dezelfde school elk op heel andere dingen letten en op grond daarvan heel verschillende beoordelingen geven. In feite is dit vooral een vaak voorkomende situatie op scholen waar elke docent een eigen subjectief vakconcept hanteert, een eigen programma volgt en daar meer of minder bewust op bouwt bij het beoordelen. Er zit dan een subjectief element in de meetprocedure (twee observatoren ‘zien’ gewoon verschillende dingen). Ook kan dit een probleem zijn als je leerlingen zelf laat meten en zij niet precies weten hoe het moet. Dan is er een gerede kans dat er meetfouten worden gemaakt. Dit pleit er dus voor om beoordelingscriteria zo veel mogelijk expliciet te maken en daarover afspraken te maken.

**D3: te Doen (hanteerbaarheid)?**

Soms moet je om het goed te doen elke leerling minstens 15 minuten in verschillende situaties aan het werk zien om een goed oordeel te kunnen geven (zie hiervoor). Maar dat betekent, als een klas uit 25 leerlingen bestaat, zo'n 25 × 15 minuten, dus ruim 6 uur beoordelen. Dat is uiteraard veel te veel. Er blijft dan geen tijd meer over om les te geven in een periode van twee tot drie weken. Dat is een serieus probleem. Je zult het dus met minder moeten doen. Aan meetmethoden mag daarom de eis worden gesteld van een zeker gebruiksgemak.

Wat gebruiksgemak betekent is een tamelijk eenvoudige zaak. Wil (systematische) opbrengstevaluatie van bewegingsonderwijs een kans krijgen, dan zullen de te gebruiken meetmethoden in elk geval eenvoudig en snel hanteerbaar moeten zijn in de lessen. Dat gebruiksgemak wordt bepaald door een aantal factoren:

* Hoeveel (les)tijd moet eraan worden besteed? Kan een bepaalde meetmethode misschien zo worden ingepast in de lessen dat zij geen extra tijd kost?
* Moeten docenten veel moeite doen om een meetinstrument ‘onder de knie’ te krijgen of kan dat instrument direct gebruikt worden?
* Moet de docent alles zelf doen of is het mogelijk om leerlingen op een eenvoudige manier te trainen in het gebruik van een meetinstrument?
* Is er veel rompslomp of ingewikkelde apparatuur mee gemoeid in de les: verschillende observatieformulieren, computers, extra organisatie, potloden en pennen? Of is er een eenvoudigere manier?
* Zit er veel administratief voor- en/of nawerk aan een meetmethode? Moeten meetresultaten eerst bewerkt worden voordat ze te interpreteren zijn?

Bij dit alles moet natuurlijk wel worden bedacht dat het gebruiksgemak altijd bekeken moet worden ten opzichte van de waarde die wordt toegekend aan de verkregen resultaten. Een meetmethode die nogal ingewikkeld en tijdrovend is maar wel valide en betrouwbare resultaten oplevert is te verkiezen boven een methode die simpel is maar waar je weinig aan hebt. Het liefst hebben we natuurlijk methoden die aan alle eisen voldoen; meestal echter zal het een kwestie zijn van afwegen en kiezen (en concessies doen). Mogelijke oplossingen kun je zoeken in efficiëntie (niet de koprol beoordelen als je ook de zweefrol nog moet doen), specificiteit (geen dingen beoordelen die niet de kern vormden van de doelen van een les of lessenreeks), leerlingen inschakelen bij het verzamelen van de informatie en de metingen efficiënt organiseren. Verder is het een kwestie van wennen aan het feit dat evalueren gewoon tijd kost en jezelf er in te trainen en er handig in worden. In de praktijk sluit je soms een compromis met je geweten tussen hoe het eigenlijk zou moeten en hoe je het feitelijk doet.

**D4: Doorzichtig (transparantie)**

Een laatste eis die aan meetmethoden gesteld kan worden is de eis van transparantie. Transparantie wil zeggen dat degene die wordt beoordeeld begrijpt waarop hij beoordeeld wordt en dat hij weet wat hij moet doen om een bepaald oordeel te krijgen. Dat is in de praktijk een kwestie van openheid tegenover de leerlingen. Deze eis komt voort uit het idee dat leerlingen er recht op hebben om te weten wat er van ze gevraagd wordt, en dat als een beoordelingsprocedure valide is (datgene vertegenwoordigt wat geëist wordt), transparantie van de beoordeling geen probleem kan zijn. Kortom, leerlingen moeten weten wat er van hen verwacht wordt omdat je anders het risico loopt dat je niet te zien krijgt wat je eigenlijk wilde zien.

Weer naar een voorbeeld. Een leerling die bij badminton elke kans afsmasht geeft jou niet de kans om te zien of hij nu eigenlijk wel of niet een tegenstander kan uitspelen door kort en lang spelen te variëren. Misschien kan hij het wel, maar hij doet het niet, waarschijnlijk omdat hij denkt dat hij wordt beoordeeld op het feit of hij scoort of niet, of hij meer punten haalt dan zijn tegenstander. Daar heb je dus net niks aan. Daarom moet altijd volstrekt helder zijn wat er precies van leerlingen gevraagd wordt. Transparantie is eigenlijk in die zin het belangrijkste criterium voor goed evalueren, omdat helderheid over de doelen en over de wijze waarop er beoordeeld zal worden het gedrag van de leerlingen stuurt. Dan heb je al bij voorbaat datgene dat je aan het eind van leerlingen wilt vragen.

Deze vier criteria gelden in elke situatie, maar niet in elke situatie op dezelfde manier. Afhankelijk van doel en belang van de evaluatie kan de ene keer validiteit wat belangrijker zijn en de andere keer de betrouwbaarheid. Je probeert wel altijd aan alle criteria te voldoen, maar leerkrachten (het zijn ook maar mensen) ontwikkelen bepaalde voorkeuren. De één houdt van helder en lekker concreet en de ander meer van open met meer flexibiliteit. Wij hebben liever dat het geen vastgeroeste voorkeuren worden, maar situatiegebonden eisen.

**Meten en evalueren ‘revisited’**

Nog één slotopmerking over meetmethoden. Het kan niet genoeg benadrukt worden dat er een principieel onderscheid gemaakt moet worden tussen meten en waarderen als elementen van evalueren. Je meet wat een leerling kan en daarvoor gelden de 4 D’s, maar daarna moet je - in het licht van de bedoeling die je met de evaluatie hebt - pas een waardeoordeel vormen over de vraag of je die prestatie goed vindt of niet – of dat nu in een cijfer is of een letter of een compliment of wat dat ook – én over de vraag of die prestatie te danken is aan het genoten onderwijs. Pas dan kun je iets zeggen over de kwaliteit van het onderwijs. En dat is telkens de centrale vraag waar we het over hebben. Bij evalueren kijk je naar de kwaliteit van onderwijs. Als je heel valide en betrouwbaar vaststelt dat een leerling heel goed kan badmintonnen wil dat nog niet zeggen dat hij dat bij jou op school heeft geleerd. Die meting zegt dus wel iets over de leerling, maar niet over het onderwijs. Om aan te tonen dat die vaardigheid het resultaat is van het gegeven onderwijs is een zwaardere bewijslast nodig.

De laatste jaren wordt er in toenemende mate gewerkt met (leerling)volgsystemen en met het bijhouden van een dossier of een portfolio. Het is goed om te bedenken dat zulke vormen van beoordelen weliswaar voortkomen uit de behoefte om meer individugericht te beoordelen (met als functie begeleiden), maar dat de uitkomsten van de 'metingen' van leerlingen net zo makkelijk selectief gebruikt kunnen worden. Dat gevaar dreigt als de stappen in een volgsysteem tot norm voor alle leerlingen worden verheven, in plaats van als kwaliteitscriteria voor de docent ter verbetering van het eigen onderwijs.