20

Kenmerken en kwaliteiten van lesobservatie-instrumenten

Marjoleine Dobbelaer
Onderzoeker, HAN University of Applied Sciences

Adrie Visscher
Hoogleraar, Universiteit Twente

Introductie

De kwaliteit van de leraar is van groot belang voor het leren van leerlingen. Het nauwkeurig meten van deze kwaliteit daarmee ook. Op die manier kunnen zoveel mogelijk leraren van voldoende niveau voor de klas komen te staan, kunnen leraren leren op welke punten zij zichzelf nog kunnen verbeteren en is het mogelijk om tijdig HRM-beslissingen te nemen als de kwaliteit van de leraar beneden het gewenste niveau is.

Het meten van lerarenkwaliteit kan op diverse manieren. Te denken valt aan het meten van de leerlingenresultaten die leraren realiseren, het meten van de leerlingpercepties van leskwaliteit of het observeren van lessen. Elke methode heeft voor- en nadelen.

In dit hoofdstuk focussen we ons op lesobservaties en de kwaliteiten en overige relevante kenmerken van de instrumenten die daarbij gebruikt worden. Een lesobservatie-instrument kent in onze definitie drie componenten (Bell et al., 2018):

- De scoringsinstrumenten met de leraargedragingen die gescoord worden.
- De maatregelen die getroffen worden met het oog op de betrouwbaarheid van de observaties, zoals de training van observatoren en de beschikbaarheid van een handleiding.
- De specificaties met betrekking tot de steekproef, zoals de kenmerken van de lessen die minimaal geobserveerd moeten worden en het aantal lessen dat minimaal geobserveerd moet worden om de gewenste uitspraken over de kwaliteit van leraren te kunnen doen.

Recent onderzoek laat zien dat het verkrijgen van betrouwbare en valide scores met een lesobservatie-instrument niet vanzelfsprekend is (bijvoorbeeld Nava et al., 2018). Aandacht voor alle drie de componenten van een lesobservatie-instrument zijn daarvoor cruciaal. In de onderwijspraktijk en het onderwijsonderzoek is er lang niet altijd aandacht voor deze drie zaken; er blijkt een grote kloof te bestaan tussen wat we weten over lesobservaties vanuit wetenschappelijk onderzoek en wat daarvan gebruikt wordt in de praktijk.

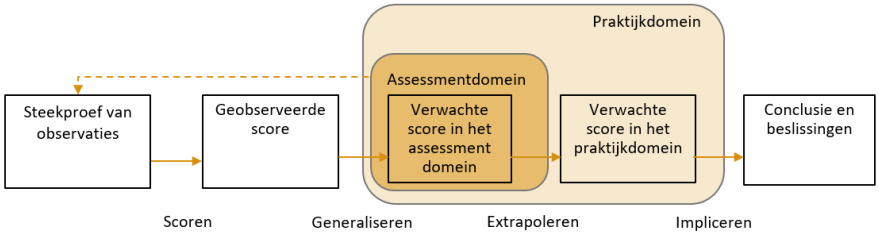
In dit hoofdstuk gaan we in op drie studies naar het gebruik van lesobservatie-instrumenten. In de eerste studie komt de vraag aan bod aan welke kwaliteitskenmerken een lesobservatie-instrument moet voldoen. De tweede studie geeft aan wat de kwaliteit is van de bestaande Nederlandse en Engelstalige lesobservatie-instrumenten voor het primair onderwijs.

Tot slot wordt in de derde studie onderzocht of externe lesobservatoren, leraren en leerlingen dezelfde les hetzelfde beoordelen. Het hoofdstuk eindigt met aanbevelingen voor ontwikkelaars en gebruikers van lesobservatie-instrumenten en de overheid.

Kwaliteitskenmerken waaraan een lesobservatie-instrument moet voldoen

Een les of leraar observeren met een observatie-instrument, maakt het mogelijk om op basis van die observatie(s) conclusies te trekken en later mogelijk beslissingen te nemen. Om te evalueren of die conclusies of beslissingen valide zijn, kan gebruikgemaakt worden van de argumentatieve benadering van validiteit (zie figuur 1). Helemaal links in de figuur staan de gedane observaties (de steekproef) en helemaal rechts de conclusies en beslissingen. De blokken ertussen representeren de redeneerlijn die over het algemeen wordt gevolgd om van de observatie(s) tot de conclusies/beslissingen te komen. Daarin zijn vier stappen te onderscheiden: scoren, generaliseren, extrapoleren en impliceren.

Figuur 1 Argumentatieve benadering van validiteit (gebaseerd op Bell et al., 2012; Kane, 2006; Wools et al., 2010).



Met de validiteitsargumentie wordt onderzocht of deze vier stappen valide zijn om in een bepaalde situatie op basis van observaties naar conclusies en beslissingen toe te werken (Kane, 2006). In de eerste stap, het scoren, wordt onderzocht of er op basis van de observatie(s) voldoende bewijs is voor het toekennen van een score. Daarna volgt de analyse of er voldoende bewijs is voor het generaliseren van de geobserveerde score. Kun je bijvoorbeeld op basis van de observatie(s) iets beweren over de didactische kwaliteit van een leraar in algemene zin (score in het assessmentdomein)? Ten derde wordt onderzocht of er voldoende bewijs is voor het extrapoleren van een score. Dit is alleen relevant als je op basis van de observatie uitspraken doet over een breder domein (het praktijkdomein) dan daadwerkelijk is geobserveerd. Een voorbeeld van extrapoleren is dat er alleen geobserveerd is bij rekenlessen, maar dat je uitspraken doet over de kwaliteit van een leraar voor alle vakken. Tot slot wordt het bewijs onderzocht voor de uitspraken die gedaan worden op basis van de observatiescores (impliceren) en of bijvoorbeeld de beslissingen passend zijn.

In de literatuur over lesobservatie(instrumenten), algemene literatuur over tests en assessments en de literatuur over validiteit komen diverse onderwerpen naar voren die van belang zijn om te adresseren bij het scoren, generaliseren, extrapoleren en impliceren met een lesobservatie-instrument. Deze onderwerpen hebben we samengevat in een evaluatiekader (tabel 1). In de tabel staat in het kort wat we verstaan onder deze onderwerpen. Daarnaast geven we per onderwerp de belangrijkste vraag weer die je als instrumentontwikkelaar of gebruiker van een observatie-instrument kunt stellen.

Tabel 1 Evaluatiekader lesobservatie-instrumenten (gebaseerd op Dobbelaer, 2019).

Onderwerp	Kwaliteitscriteria observatie-instrument	Belangrijke vraag om te stellen als instrument-ontwikkelaar	Belangrijke vraag om te stellen als gebruiker
Scoren			
Bedoeld gebruik	Duidelijke omschrijving van: • Het construct dat gemeten wordt • De context waarvoor het ontwikkeld is • Het doel van het instrument	Is het bedoelde gebruik voor (potentiële) gebruikers duidelijk omschreven, zodat het voor hen duidelijk is of dat past binnen de eigen context?	Past het bedoelde gebruik van het instrument bij mij, zo niet, welke implicaties heeft dit? Bijvoorbeeld als een instrument voor formatief gebruik wordt ingezet voor een summatief doel.
Kwaliteit van de items in het instrument	• De wetenschappelijke basis van de items • De inhoudsvaliditeit en begripsvaliditeit • De kwaliteitseisen van de items	Is het voor (potentiële) gebruikers duidelijk waar de items op zijn gebaseerd en hoe de items tot stand zijn gekomen (operationalisatie)?	Is de inhoudsvaliditeit en constructvaliditeit bewezen?
Kwaliteit van de scoringsregels	• Passende regels die inzichtelijk maken wanneer een bepaalde score wordt toegekend • Passende regels om van itemscores tot een overall score te komen (bijvoorbeeld het gemiddelde)	Worden de scoringregels voldoende (empirisch) ondersteund?	Worden de scoringregels ondersteund in mijn eigen context? Bijvoorbeeld: vinden wij lessen die met dit instrument als voldoende worden aangemerkt, op school ook voldoende?
Maatregelen voor observatoren	• Maatregelen die ervoor zorgen dat verschillen tussen observatoren verkleind worden, zoals een handleiding, training of scoringsregels	Wat bied ik gebruikers om de kwaliteit van observatoren te waarborgen?	Wie zijn mijn observatoren, en hoe zorg ik ervoor dat zij daadwerkelijk meten wat ik wil meten en dat de observatie niet te veel afhankelijk is van één persoon?
Standaardisatie van observatie-procedures	• Richtlijnen voor de observatie om verschillen tussen observatoren te verkleinen, zoals de mate waarin observatoren tijdens de observatie interactie mogen hebben met kinderen	Welke richtlijnen geef ik om observaties te standaardiseren?	Welke richtlijnen worden bij het instrument omschreven en welke gevolgen heeft dit voor de validiteit van de score als ik hiervan afwijk?
Bewijs voor betrouwbaar instrument-gebruik	• Empirisch bewijs voor de betrouwbaarheid van de items en de schaal	Welk empirisch bewijs heb ik voor de kwaliteit van het instrument en de betrouwbaarheid van observatoren?	Welk bewijs voor betrouwbaar instrumentgebruik kan ik in mijn eigen context verzamelen? Bijvoorbeeld over de betrouwbaarheid van de observatoren.
Keuze voor (aantal) observatoren	• Empirisch bewijs voor de interbeoordelaarsbetrouw-baarheid	Kan ik empirisch bewijs verzamelen voor het benodigde aantal observatoren?	Wie kan het beste bepaalde lessen observeren, kijkend naar de te observeren les en leraar? Bijvoorbeeld niet elke keer dezelfde observator of geen bekende van de leraar.

Onderwerp	Kwaliteitscriteria observatie-instrument	Belangrijke vraag om te stellen als instrument-ontwikkelaar	Belangrijke vraag om te stellen als gebruiker
Generaliseren			
Representa-tiviteit van de steekproef	• Het instrument bevat richtlijnen voor de keuze van de te observeren lessen (bijvoorbeeld het type lessen), wanneer geobserveerd moet worden (gedurende het jaar, de week, de dag) en voor de vereiste duur van de observatie.	Wat moeten gebruikers observeren om een goed beeld van het assessmentdomein te krijgen?	Waarover wil ik precies uitspraken doen en wat betekent dit voor mijn steekproef? Bijvoorbeeld als ik een uitspraak wil doen over het functioneren in een heel schooljaar, wanneer plan ik dan de observatie(s)?
Steekproef-grootte	• Het instrument bevat richtlijnen voor het aantal te observeren lessen bij bepaalde doeleinden, bijvoorbeeld vier lessen, als het gaat om een summatieve beoordeling.	Kan ik empirisch bewijs verzamelen voor het benodigde aantal lessen?	Zijn de richtlijnen bij het observatie-instrument haalbaar en wat betekenen die voor de implicaties als mijn steekproef hier niet aan voldoet? Bijvoorbeeld als ik één les observeer in plaats van vier.
Extrapoleren			
Relatie tussen het assessment-domein en het praktijkdomein	• Het assessmentdomein lijkt een groot deel van het praktijkdomein te dekken ('face validity'). • De theoretische basis van het instrument laat zien dat de inhoud van het lesobservatie-instrument past binnen het praktijkdomein. • De score op het assessmentdomein hangt samen met een andere meting binnen het praktijkdomein.	Kan ik empirisch bewijs verzamelen voor de relatie tussen het assessmentdomein en het praktijkdomein?	Is het aannemelijk dat ik met de score in het assessmentdomein ook iets kan zeggen over het praktijkdomein?
Impliceren			
Geschiktheid van het voorgestelde gebruik en implicaties	• De implicaties die worden aangegeven op basis van het eindoordeel in het praktijkdomein zijn terecht.	Kan ik gebruikers richtlijnen geven voor welke implicaties passend kunnen zijn voor het instrument en onder welke voorwaarden?	Passen de conclusies en beslissingen bij de meting die ik verricht heb? Zowel inhoudelijk als wat betreft de kenmerken van mijn meting (bijvoorbeeld qua aantallen observaties)?

Het evaluatiekader is relevant voor instrumentontwikkelaars, omdat het aangeeft welke aspecten van belang zijn bij de ontwikkeling van een lesobservatie-instrument. Het is ook bedoeld voor (potentiële) gebruikers van bestaande instrumenten, namelijk om de kwaliteit van deze instrumenten en het gebruik van een instrument in de eigen context te evalueren. Zelf hebben we het evaluatiekader gebruikt om bestaande lesobservatie-instrumenten voor het primair onderwijs te beoordelen.

De kwaliteit van bestaande lesobservatie-instrumenten voor het primair onderwijs

Diverse lesobservatie-instrumenten zijn ontwikkeld door onderzoekers, mensen in de praktijk, overheden en commerciële partijen. We zijn op zoek gegaan naar Nederlandse en Engelse instrumenten die ontwikkeld zijn om (een aspect van) leraarkwaliteit in het primair onderwijs te meten, om deze instrumenten vervolgens te beoordelen met het gepresenteerde evaluatiekader (Dobbelaer, 2019). Een voorwaarde voor de beoordeling van een lesobservatie-instrument was dat er onderzoek naar het instrument beschikbaar moest zijn. Opvallend was dat de meeste lesobservatie-instrumenten die beschikbaar zijn voor het primair onderwijs in Nederland hierdoor afvielen. Deze instrumenten werden ten tijde van onze zoektocht wel aangeboden/verkocht aan scholen, maar er was geen enkel onderzoek dat aantoonde dat de instrumenten betrouwbare/valide scores van leraarkwalitiet zouden kunnen opleveren.

In totaal voldeden 27 lesobservatie-instrumenten aan onze zoekcriteria. Elk instrument is met het evaluatiekader beoordeeld door twee getrainde beoordelaars. De resultaten laten zien dat de beoordelaars van de drie componenten in onze definitie van lesobservatie-instrumenten het meest positief waren over de scoringsinstrumenten. In de meeste instrumenten was ook aandacht aan de betrouwbaarheid van de observatoren (de tweede component). Echter, bij de meeste instrumenten betwijfelde ten minste één beoordelaar of de maatregelen ervoor zouden zorgen dat observatoren met het instrument accuraat en betrouwbaar kunnen scoren. Het minst positief waren de beoordelaars over de derde component: instrumentontwikkelaars waren volgens hen weinig expliciet over de specificaties met betrekking tot de steekproef. Voor de meeste lesobservatie-instrumenten gold ten slotte ook dat de beoordelaars vaak niet overtuigd waren van het bewijs voor betrouwbare en valide scores. Dit kwam vooral door het ontbreken van gedegen onderzoek naar de instrumenten.

Deze resultaten geven geen positief beeld van de kwaliteit van de lesobservatie-instrumenten. Instrumentontwikkelaars lijken vooral te focussen op de ontwikkeling van scoringsinstrumenten. Mogelijk zijn er onvoldoende middelen (geld en tijd) om ook aandacht te besteden aan de twee andere componenten in onze definitie en om gedegen onderzoek naar het instrument te doen. Een andere mogelijke verklaring is dat instrumentontwikkelaars en -gebruikers zich onvoldoende bewust zijn van het belang hiervan.
Bovenstaande laat zien dat het verkrijgen van valide lesobservatiescores geen gemakkelijke taak is. We hebben daarom ook onderzocht in hoeverre observatiescores samenhangen met twee andere metingen van leraarkwaliteit.

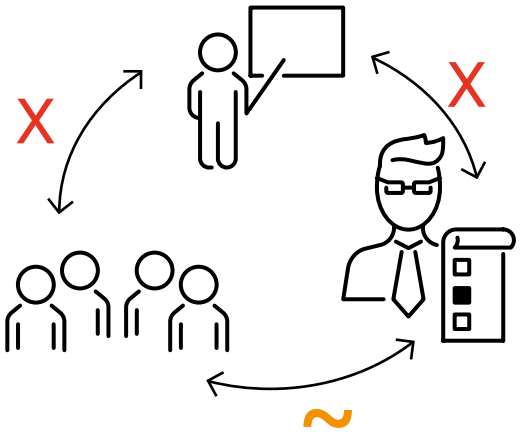
Beoordelen externe observatoren, leraren en leerlingen de leraarkwaliteit hetzelfde?

Het is nog erg onduidelijk hoe de resultaten van verschillende methoden om leraarkwaliteit te meten met elkaar samenhangen. In deze studie hebben we onderzocht hoe de beoordelingen van een les door getrainde observatoren overeenkomen met de percepties van leerlingen en leraren over dezelfde les (Dobbelaer, 2019). Hiervoor zijn data verzameld onder 25 wiskundeleraren en hun 608 havo 3-leerlingen. Voor alle leraren hebben we data verzameld over drie lessen.

De percepties van de leerlingen over de leskwaliteit van hun leraar is gemeten met de Impact!-tool. Dit is een applicatie waarmee leerlingen aan het einde van een les de leraar feedback kunnen geven over de kwaliteit van de gegeven les. De leerlingen beantwoorden in deze applicatie vijftien gesloten vragen op een vierpuntsschaal. Deze vragen zijn gebaseerd op wetenschappelijke literatuur over kenmerken van effectieve lessen, bijvoorbeeld of de uitleg door de leraar voor de leerlingen begrijpelijk is. In ons onderzoek hebben de leraren dezelfde vijftien vragen beantwoord aan het einde van de les. Alle 75 lessen zijn ook opgenomen op video en achteraf beoordeeld door drie getrainde observatoren op basis van dezelfde vijftien items.

De leraren waren in dit onderzoek gemiddeld het meest positief over hun leskwaliteit, gevolgd door hun leerlingen. De externe observatoren waren gemiddeld een stuk minder positief. De beoordelingen door de leerlingen en door de leraren hingen echter helemaal niet met elkaar samen, zoals te zien is in figuur 2. Hetzelfde gold voor de beoordelingen door de externe observatoren en die door leraren. We vonden een matige samenhang tussen de beoordelingen door de externe observatoren en die door de leerlingen.

Figuur 2 Mate van overeenstemming tussen observatoren, leerlingen en leraren.



Mogelijke verklaringen voor deze resultaten:

- Observatoren, leerlingen en leraren hebben een andere visie gehad op bepaalde aspecten. Bijvoorbeeld wat moet het benoemen van een lesdoel inhouden voor een voldoende score? Of wanneer is er hard genoeg gewerkt tijdens een les voor een voldoende score?
- De scores zijn beïnvloed door andere factoren. Bijvoorbeeld leraren hebben zichzelf hoger beoordeeld, omdat ze wisten dat onderzoekers naar de antwoorden zouden kijken (sociale wenselijkheid). Of leerlingen hebben de les anders beoordeeld door de invloed van andere factoren, zoals de cijfers die ze net gekregen hebben.
- Het is moeilijk voor externe observatoren en leraren om het effect van lesaspecten op leerlingen te beoordelen. Van de vijftien vragen die de beoordelaars hebben beantwoord, gingen er diverse over de beleving van leerlingen. Bijvoorbeeld of ze zich veilig voelden in de klas en of de leraar vragen heeft gesteld die hen aan het denken zetten. Het kan zijn dat dergelijke vragen moeilijk te observeren zijn en het beste aan leerlingen gesteld kunnen worden.

De resultaten laten zien dat het bij de beoordeling van leraarkwaliteit belangrijk is om goed na te denken over welke meetmethode in een bepaalde situatie het meest geschikt is, omdat elke methode kan leiden tot een ander resultaat en mogelijk niet alle aspecten van leskwaliteit met elke methode het beste gemeten kunnen worden.

Conclusie en aanbevelingen

Op basis van de resultaten uit de studies kunnen we een aantal aanbevelingen doen voor instrumentontwikkelaars, gebruikers van lesobservatie-instrumenten en de overheid.

Ontwikkelaars van lesobservatie-instrumenten

De review laat zien dat de meeste lesobservatie-instrumenten die ten tijde van de review werden gebruikt, in de praktijk en in onderzoek niet aan de standaarden voldeden die in het evaluatiekader zijn omschreven. Om ervoor te zorgen dat lesobservaties valide scores opleveren, zouden instrumentontwikkelaars meer aandacht moeten krijgen voor:

- alle drie de componenten van onze definitie van lesobservatie-instrumenten;
- onderzoek naar de betrouwbaarheid en validiteit van de scores met de instrumenten;
- het geven van duidelijke richtlijnen voor het gebruik van de instrumenten op basis van dat onderzoek.

Instrumentontwikkelaars kunnen op die manier de schakel vormen tussen wetenschappelijk onderzoek en het gebruik van lesobservatie-instrumenten in de praktijk. Daarvoor is het nodig dat zij zich bewust zijn van de waarde van de kwaliteitsstandaarden in het evaluatiekader en de complexiteit van de ontwikkeling van een lesobservatie-instrument.

Gebruikers van lesobservatie-instrumenten

Het is belangrijk dat ook de gebruikers van lesobservatie-instrumenten op de hoogte zijn van de kwaliteitsstandaarden in het evaluatiekader en het belang ervan inzien. Hierdoor kiezen zij hopelijk alleen nog instrumenten die aan de standaarden voldoen en houden zij zich aan de richtlijnen die instrumentontwikkelaars opstellen voor het gebruik ervan op basis van onderzoek. Dit kan het verkrijgen van observatiescores duurder en tijdsintensiever maken, bijvoorbeeld omdat observatoren getraind moeten worden of doordat meerdere observaties nodig zijn. Het is echter essentieel dat gebruikers van lesobservatie-instrumenten inzien dat het verkrijgen van onbetrouwbare of niet-valide scores geen waarde heeft en een verspilling van middelen is.

Het is ook van belang dat gebruikers van observatie-instrumenten goed kijken naar het instrumentgebruik in hun eigen context. Deze context kan de betrouwbaarheid en validiteit van de scores beïnvloeden, bijvoorbeeld als

leraren worden geobserveerd door observatoren die voor hen bekend zijn. De vragen bij het evaluatiekader in tabel 1 kunnen hierbij helpen.

Op basis van de resultaten in de derde studie kan het voor gebruikers van lesobservatie-instrumenten ook relevant zijn om lesobservaties met andere meetmethoden te combineren.

Overheid

De overheid kan op verschillende manieren het gebruik van lesobservatie-instrumenten die aan de kwaliteitsstandaarden voldoen, stimuleren:

- Door een beoordelingssysteem voor lesobservatie-instrumenten te ontwikkelen.
- Door onafhankelijke beoordelingen van de lesobservatie-instrumenten, zoals de COTAN doet voor de kwaliteit van testen, mogelijk te maken.
- Door een database met alleen instrumenten die aan de kwaliteitsstandaarden voldoen, te ontwikkelen.

Als duidelijk is welke instrumenten aan de kwaliteitscriteria uit het evaluatiekader voldoen, kan de overheid van scholen verlangen om alleen middelen te besteden aan het gebruik van dergelijke instrumenten.

Literatuur

Bell, C., Dobbelaer, M. J., Klette, K., & Visscher, A. J. (2018). *Qualities of classroom observation systems. School Effectiveness and School Improvement.* doi: 10.1080/09243453.2018.1539014

Bell, C. A., Gitomer, D. H., McCaffrey, D. F., Hamre, B. K., Pianta, R. C., & Qi, Y. (2012). *An argument approach to observation protocol validity. Educational Assessment, 17*(2–3), 62-87. doi: 10.1080/10627197.2012.715014

Dobbelaer, M. J. (2019). *The quality and qualities of classroom observation systems*. Geraadpleegd op https://research.utwente.nl/en/publications/the-quality-and-qualities-of-classroom-observation-systems

Kane, M. T. (2006). Validation. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 17-64). Westport,CT: American Council on Education and Praeger Publishers.

Nava, I., Park, J., Dockterman, D., Kawasaki, J., Schweig, J., Quartz, K. H., & Martinez, J. F. (2018). Measuring teaching quality of secondary mathematics and science residents: A classroom observation framework. *Journal of Teacher Education*. doi: 10.1177/0022487118755699

Wools, S., Eggen, T., & Sanders, P. (2010). Evaluation of validity and validation by means of the argument-based approach. *CADMO, 8,* 63-82.