Hoe zijn competenties grootschalig te toetsen? Ontwikkeling van een evaluatiematrix voor toetsprogramma's en een inventarisatie van 'good practices'.

Beleidssamenvatting

Promotor-woordvoerder:
Prof. dr. Sven De Maeyer
Promotoren:
Prof. dr. Vincent Donche
Prof. dr. Jan Vanhoof
Prof. dr. Peter Van Petegem
Projectmedewerkers:
Dr. Liesje Coertjens
Dr. Jetje De Groof
Mevr. Alexia Deneire

Vlaams Ministerie van Onderwijs en Vorming

1 Inleiding

Competentiegericht onderwijs maakt de omslag naar competentiegericht beoordelen en evalueren noodzakelijk. Het meten van brede, complexe constructen zoals competenties vereist het inschakelen van performance assessment. Het kwaliteitsvol opzetten van performance assessment is echter geen eenvoudige taak, zeker niet wanneer de toetsing grootschalig is, zoals in het geval van toetsen gericht op systeemmonitoring.

Met het oog op het voorzien van een empirische fundering voor de stap naar meer competentiegericht peilingsonderzoek, heeft deze studie als focus: (1) een stand van zaken te geven van de inzichten m.b.t. de kwaliteitseisen van performance assessment; (2) op basis van deze kwaliteitseisen een evaluatiematrix uit te werken om toetsprogramma's op basis van hun theoretische en praktische sterktes en zwaktes te positioneren; (3) op basis van de evaluatiematrix buitenlandse voorbeelden van grootschalige competentiebeoordelingen te inventariseren en te duiden. Vanuit deze doelstellingen worden vier resultaten vooropgesteld die richtinggevend zijn voor de rest van het onderzoek:

1. Ontwikkeling van een evaluatiematrix die grootschalige performance assessments, gericht op kwaliteitsbewaking op systeemniveau, op hun kwaliteit toetst. Deze ontwikkeling houdt de identificatie in van de noodzakelijke bouwstenen.
2. Inventarisatie en evaluatie van recente empirische literatuur over kwaliteitseisen die aan performance assessments gesteld worden.
3. Inventarisatie en evaluatie van internationale praktijkvoorbeelden van grootschalige competentietoetsen die monitoring op systeemniveau als doelstelling hebben.
4. Identificatie van uitdagingen en oplossingen bij grootschalige performance assessments die monitoring op systeemniveau als doelstelling hebben.

De kwaliteitseisen die gesteld worden aan grootschalige competentiebeoordelingen via performance assessment, vormen de rode draad van deze doelstellingen. Centraal staan de begrippen ‘competentie’, ‘performance assessment’, ‘kwaliteit’ en ‘monitoring op systeemniveau’. De volgende werkdefinities die voor deze begrippen werden uitgewerkt, zorgen voor de nodige afbakening en focus van de voorliggende studie.

*Een **competentie** verwijst naar de bekwaamheid om specifieke combinaties van kennis, vaardigheden en attitudes in te zetten bij het volbrengen van een specifieke taak, relevant voor persoonlijke, professionele of maatschappelijke activiteiten*.

***Performance(-based) assessment** betreft beoordeling (van competenties) waarbij gebruik wordt gemaakt van (levensechte) taken, relevant voor de beoogde competenties.*

***Kwaliteit** wordt in deze studie opgevat als een combinatie van psychometrische elementen zoals validiteit en betrouwbaarheid en 'alternatieve' criteria zoals authenticiteit, transparantie en eerlijkheid. Deze verschillende kwaliteitscriteria worden voortdurend tegen elkaar afgewogen, waarbij ook gekeken wordt naar de haalbaarheid van de opzet van de toets in termen van tijd, financiële middelen en infrastructuur.*

*Toetsen die **monitoring op systeemniveau** beogen zijn grootschalige toetsen die rapporteren over wat groepen van leerlingen kennen en kunnen, in relatie tot vooraf vastgelegde onderwijsdoelstellingen. Omdat de resultaten worden gerapporteerd op systeemniveau, hebben ze geen repercussies voor individuele leerlingen, en worden ze als 'low-stakes'-toetsen beschouwd.*

Samenvattend komt de studie neer op een zoektocht naar kwaliteitsindicatoren van toetsen, die gebruik maken van levensechte taken en die daarbij meer toetsen dan geïsoleerde kennis of vaardigheden. Daarbij wordt kwaliteit opgevat als een combinatie van psychometrische en alternatieve criteria. De interesse gaat primair uit naar grootschalige toetsen die monitoring op systeemniveau beogen.

2 Theoretisch kader

In wat volgt verduidelijken we eerst waarom we de argumentatieve benadering van validiteit (Kane, 2006, 2013; Kane et al., 1999) als uitgangspunt hebben gekozen, om vervolgens de kernelementen van de visie van deze benadering aan te stippen. Daarna zoeken we bij andere auteurs aanknopingspunten die ons kunnen helpen 1) deze argumentatieve benadering te gebruiken voor de evaluatiematrix die we willen ontwikkelen, en 2) extra kwaliteitsbouwstenen van performance assessment te identificeren waar de argumentatieve benadering op zich minder expliciet aandacht aan besteedt. Op die manier komen we tot een selectie van bouwstenen die noodzakelijk zijn om in het kader van grootschalige competentietoetsing kwaliteitsvolle performance assessments te kunnen opzetten.

2.1 Argumentatieve benadering van validiteit

Om de kwaliteit van toetsen (inclusief meer omvattende toetsprogramma's) te evalueren zijn er verschillende kaders ontwikkeld (bv. Baartman, 2008; Evers, Lucassen, Meijer, & Sijtsma, 2009; Kane, 2006, 2013; Wools, 2015; Wools et al., 2007). Vele auteurs (bv. Brennan, 2006; Chapelle, 2012; Chapelle et al., 2010; Crisp & Shaw, 2011; Enright & Quinlan, 2010; Llosa, 2008; Schuwirth & van der Vleuten, 2012; Shaw et al., 2011; Wools, 2012, 2015) schuiven de argumentatieve benadering van validiteit (Kane, 2006, 2013; Kane et al., 1999) naar voor als een generiek toepasbaar en praktisch model om de kwaliteit (i.c. validiteit) van toetsen en beoordelingssystemen in kaart te brengen.

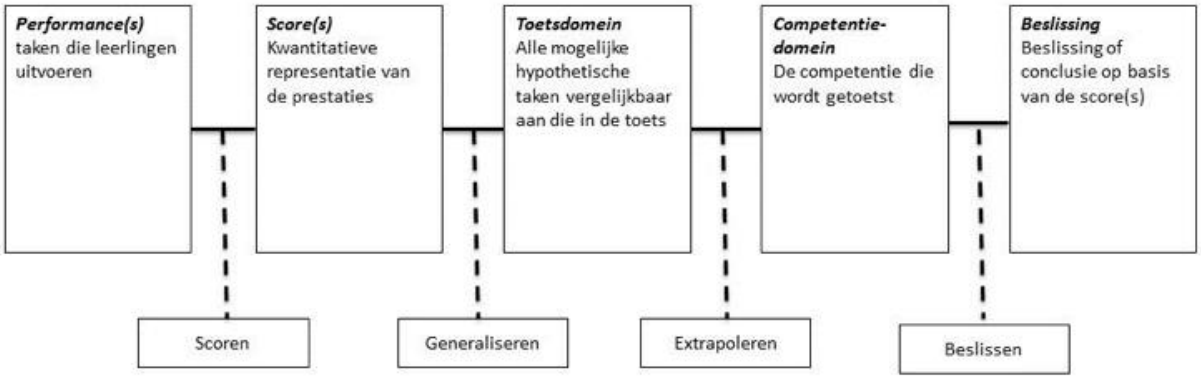
De argumentatieve benadering vormt, in vergelijking met andere invalshoeken, een hanteerbare leidraad om de validiteit van interpretaties en gebruik van toetsscores na te gaan. Enerzijds worden er duidelijke stappen voorgesteld, die onontbeerlijk zijn bij het verzamelen van validiteitsbewijzen. Anderzijds biedt Kane's aanpak ook handvaten om de argumentatie van validiteit op een methodologisch solide basis te onderbouwen.

De centrale vraag in elk valideringsvraagstuk is hoe we op grond van prestaties op een beperkte set van toetstaken uit het toetsdomein, tot valide conclusies kunnen komen omtrent verwachte 'performance' in het beoogde competentiedomein. Kane (2006, 2013) stelt voor om dit probleem in kleine deelproblemen op te splitsen en te werken in stappen.

De vertaalslag tussen prestaties aan de ene zijde van de keten en een uiteindelijke beslissing op basis van de toetsscore, inclusief de gevolgen van de interpretatie, aan de andere zijde van de keten, omvat volgens Kane (2006, 2013) steeds vier te onderscheiden stappen:

1. scoren of het vertalen van geobserveerde prestaties in scores;
2. generaliseren van de toegekende scores naar scores voor een welbepaald toetsdomein ('universe of generalization');
3. extrapoleren van deze scores voor het toetsdomein naar het beoogde competentiedomein;
4. het nemen van beslissingen.

Figuur 1 visualiseert deze keten, die door Kane zelf het 'interpretatie- en gebruiksargument' (Kane, 2006, 2013) wordt genoemd: een logische argumentatie om vanuit de observatie van toetsprestaties te (kunnen) veralgemenen naar de vaardigheid of competentie waarin men initieel is geïnteresseerd en waarvoor men het assessment heeft opgezet.



Figuur 1: Argumentatief model van validiteit.

Samengevat betekent dit dat de prestatie wordt beoordeeld, wat leidt tot een geobserveerde score. Indien aangetoond kan worden dat de geobserveerde score representatief is voor de (hypothetische) score op alle mogelijke taken, verkregen op alle mogelijke afnamemomenten, door alle mogelijke beoordelaars, kan de score gegeneraliseerd worden naar het volledige toetsdomein (het 'universum'). Om vervolgens de verwachte score voor het toetsdomein te kunnen extrapoleren naar het ruimere competentiedomein, moet aangetoond worden dat de toetstaken adequate maten bieden voor de competentie of het construct waarin men is geïnteresseerd en dat de prestaties op de taken een goede indicator zijn voor prestaties op (criterium)taken in het echte leven. Tot slot wordt een beslissing genomen omtrent het al dan niet halen van de prestatiestandaard(en) met betrekking tot de beoogde competentie, wat betekent dat men verdere consequenties aan de score verbindt (bv. bepaalt of de betreffende leerlingenpopulatie een welbepaalde eindterm al of niet haalt). Telkens er toetsresultaten gebruikt worden om conclusies te trekken of beslissingen te nemen wordt deze logische argumentatie toegepast (Kane, 2006).

Met betrekking tot de stappen generaliseren en extrapoleren wijst Kane (Kane et al., 1999, Kane, 2013), zeker in het geval van performance assessments, op een onvermijdelijke paradox. Deze paradox houdt verband met het delicate evenwicht tussen de betreffende stappen. Algemeen gesteld namelijk ondersteunt standaardisering de stap van het generaliseren doordat dit het toetsdomein in zeker mate verengt; tegelijkertijd ondermijnt ze daarmee de mogelijkheid tot extrapoleren. Standaardisering leidt er namelijk toe dat het universum ten opzichte van het boogde competentiedomein verkleint.

Eigen aan deze sequentiële keten is dat we, met het oog op validering, de aanwezige deducties en onderliggende veronderstellingen in het argument expliciteren en de bewijzen aan een serie kritische tests (logische analyses en empirische studies) onderwerpen. De validiteit van (competentie)toetsscores kan dan worden geëvalueerd of aangetoond door te expliciteren hoe de verschillende stappen of gevolgtrekkingen in de opgebouwde argumentatie worden gewaarborgd door analytische of empirische evidentie (Toulmin, 1958).

Zo hangt de validiteit van de interpretatie en het daaropvolgende gebruik van de toetsscores, resulterend uit de competentiebeoordeling, af van de plausibiliteit van de deducties in het argument (Kane, 2006, 2013; Kane et al., 1999). Kane (2006, 2013) legt de klemtoon namelijk op de bedoelde interpretatie en het beoogde gebruik van toetsscores. Het duidelijk specificeren van beoogde interpretatie en gebruik is noodzakelijk om de veronderstellingen en gevolgtrekkingen die cruciaal zijn voor de interpretatie en het gebruik, te kunnen identificeren.

Belangrijk in het validiteitsargument is dat de keten maar zo sterk is als de zwakste schakel; in het geval van een (enkele) ongeldige gevolgtrekking wordt de plausibiliteit van de volledige argumentatie ondermijnd, ook wanneer het bewijs voor de overige deducties overweldigend is (Crooks, Kane, & Cohen, 1996; Kane et al., 1999). Kane et al. (1999) stellen de keten van deducties in dat opzicht voor als bruggen die aaneen geschakeld zijn en die allemaal overgestoken moeten worden; indien één van de bruggen onvoldoende sterk blijkt, valt het argument volledig in het water. Dit impliceert overigens ook dat het geen zin heeft (een) volgende stap(pen) te gaan onderzoeken indien de vorige stap niet of onvoldoende plausibel bleek.

De bewijzen voor de deducties kunnen empirisch zijn of logisch-analytisch. Het eerste type bewijzen kan verzameld worden vooral tijdens de pilotering van de toets en op grond van (statistische) analyses op de verzamelde gegevens. Analytische bewijzen (bv. verslagen over de rationale van de item- en taakconstructie) worden voornamelijk tijdens de ontwikkelingsfase van de toets gegenereerd, bij de ontwikkeling van het interpretatieve argument en de meetprocedure (Kane, 2006). De evaluatie van interpretaties van en gebruik van toetsscores (validiteitsargument) vereist overigens dat verschillende soorten bewijzen afkomstig uit diverse bronnen samen gelegd worden; een enkele analyse of studie volstaat niet, er moet sprake zijn van een serie kritische tests (Kane, 2006).

Wat ten slotte ook bijdraagt tot een correct begrip van de argumentatieve benadering van validiteit, is de noodzaak om toetsprestaties en - scores enerzijds en de interpretatie en implicaties ervan anderzijds los te koppelen van elkaar: de interpretatie en/of het gebruik van een toetsprestatie is al of niet (voldoende) valide; niet de prestatie, toetsscore of toetsprocedure an sich (Kane, 2013; Kane et al., 1999).

De notie ‘voldoende’, brengt ons op een ander fundamenteel gegeven, namelijk dat het bepalen van de validiteit geen kwestie van alles of niets is; over validiteit spreken we in termen van minder of meer valide. Op grond van de kritische evaluatie van alle inferenties, onderliggende assumpties en bewijsmateriaal doen we uitspraken over de mate van validiteit van de interpretatie die we geven aan de scores op een welbepaalde competentietoets.

2.2 Variaties en/of aanvullingen op de benadering van Kane

Bij de ontwikkeling van de evaluatiematrix consulteerden we naast Kane ook andere auteurs (Chapelle et al., 2010; Crooks et al., 1996; Shaw et al., 2011; Wools, 2015). Deze auteurs vertrekken evenzeer vanuit een argumentatieve benadering van validiteit. De keten van scoren, generaliseren, extrapoleren en beslissen vormt net als bij Kane de kern van hun betoog. De specifieke invulling van het valideringskader van deze auteurs vertoont echter interessante nuanceverschillen, die enerzijds het gevolg zijn van het vertrekken uit andere toetscontexten (bv. taaltoetsen, traditionele toetsen of meer competentiegerichte assessments); anderzijds hebben ze te maken met de nadruk die auteurs willen leggen op specifieke aspecten van het beoordelingsproces. In de volgende paragrafen schetsen we kort om welke verschilpunten het gaat en welke lessen wij daaruit trekken met het oog op de verdere uitwerking van onze evaluatiematrix. Met betrekking tot de manier waarop we dit uitwerken maken we gebruik van het stramien van het onderzoeksrapport van Curcin, Boyle, May en Raman (2014).

Crooks en collega’s

Crooks et al. (1996) stellen een kader voor waarin de chronologie en logica van het beoordelingsproces nadrukkelijker aanwezig zijn dan in de eerder besproken aanpak van Kane. Het beoordelingsproces wordt voorgesteld als een keten van de volgende aaneen geschakelde fasen: toetsafname, scoren, aggregeren, generaliseren, extrapoleren, beslissing en impact.

Interessant voor onze studie is het feit dat toetsafname en scoren uit elkaar gehaald worden. We zijn van menig dat dit onderscheid ook voor onze evaluatiematrix een belangrijke toegevoegde waarde kan bieden, in die zin dat in de fase van de toetsafname belangrijke valkuilen schuilen die, indien ze niet gemeden worden, de generalisatie en extrapolatie van de scores van de toets hypothekeren.

Praktisch bruikbaar vinden we bovendien dat in deze studie niet de volledige argumentatiestructuur wordt meegenomen, maar dat de nadruk gelegd wordt op het in kaart brengen van de belangrijkste bedreigingen, die in elke aparte fase de kop opsteken.

Chapelle en collega's

Het interpretatieve- en gebruiksargument dat Chapelle en haar collega's in hun studie voorstellen, onderscheidt zes gevolgtrekkingen met bijhorende principes of vuistregels en veronderstellingen: domeinbeschrijving, evalueren, generaliseren, uitleggen, extrapoleren en toepassen (Chapelle, 2012; Chapelle et al., 2010).

Het opnemen van de stap 'domeinbeschrijving' vloeit rechtstreeks voort uit volgende observatie van (Kane, 2004, p. 141):

> *(…) if the test is intended to be interpreted as a measure of competence in some domain, then efforts to describe the domain carefully and to develop items that reflect the domain (in terms of content, cognitive level, and freedom from potential sources of systematic errors) tend to support the intended interpretation.*

In het kader van de 'Test of English as a Foreign Language' (TOEFL) verantwoorden en expliciteren Chapelle en collega's (2010, p. 8) deze keuze als volgt:

> *The validity of that inference rests on the assumptions that assessment tasks that are representative of the academic domain can be identified, that critical English language skills, knowledge, and processes needed for study in English-medium colleges and universities can be identified, and that assessment tasks that require important skills and are representative of the academic domain can be developed.*

Net omwille van het belang van deze koppeling tussen het beoogde competentiedomein en de toetstaken neemt Chapelle 'domeinbeschrijving' expliciet op in het interpretatieve argument van de TOEFL iBT (Chapelle, 2012).

Shaw en collega's

Ook in het theoretisch kader van hun studie baseren Shaw et al. (2011) zich op het denken van Kane. Ze benadrukken de nood aan een concreet toepasbaar kader en komen in hun zoektocht bij de volgende stappen/gevolgtrekkingen terecht: constructrepresentatie, scoren, generaliseren, extrapoleren en beslissen.

In dit kader wordt aan elke gevolgtrekking die verantwoord moet worden, een valideringsvraag gelinkt. Deze aanpak vinden we voor onze evaluatiematrix een verbetering naar praktische bruikbaarheid toe. Op die manier is de matrix ook bruikbaar voor niet-methodologen. Hoewel er geen specifieke veronderstellingen worden geformuleerd die aan de basis van elke gevolgtrekking liggen, zijn deze assumpties impliciet aanwezig in de geformuleerde vragen.

Net als Chapelle et al. (2010) schuiven deze auteurs bovendien 'constructrepresentatie' als extra (eerste) stap naar voor. Hiermee onderschrijven ook zij het belang van een grondige domeinbeschrijving.

Wools

Het valideringskader van Wools (2015), ten slotte, werd ontwikkeld binnen de context van competentiebeoordelingen in het beroepsonderwijs. Het interpretatieve argument dat zij voorlegt vertaalt de uitvoering van een bepaalde taak in een beslissing aangaande iemands bekwaamheid of competentie via de volgende keten van gevolgtrekkingen: performance, scoren, toetsdomein, competentiedomein, praktijkdomein, beslissing.

Vernieuwend is hier de expliciete opdeling van de fase van het extrapoleren in twee stappen. Een eerste stap omvat de mogelijkheid tot extrapolatie van het toetsdomein naar het competentiedomein (een operationalisatie van de competentie die gemeten wordt); de volgende stap trekt de extrapolatie door van dat competentiedomein naar het prakijkdomein. Het praktijkdomein omvat dan situaties uit het dagdagelijkse leven die mensen kunnen tegenkomen in hun toekomstige beroepsleven.

De eerste stap die (Wools, 2015) onderscheidt, de extrapolatie naar het toetsdomein, bestaat uit de operationalisering van de competentie die gemeten wordt, en valt in feite dus samen met wat Chapelle domeinbeschrijving noemt. Dit heeft ermee te maken dat indien men in de ontwikkelingsfase van de toets een goede domeinbeschrijving vastlegt, waarbij ook experts en vertegenwoordigers uit het werkveld betrokken zijn, de kans op extrapolatie van de scores op de toets naar het competentie- en vervolgens ook naar het praktijkdomein, beduidend groter wordt.

Ook voor de stap ‘evalueren’ (met daaronder ‘toetsafname’ en ‘scoren’) geldt dat de bewijzen die men kan verzamelen in het kader van deze stap, te maken hebben met een kwaliteitsvol opzet van de toets tijdens de designfase en de kwaliteitscontrole tijdens de designfase van de toets. Indien deze stappen optimaal worden uitgevoerd, vergroot dit de kansen dat de scores op de toets vervolgens te generaliseren en/of te extrapoleren zijn. We denken dan bijvoorbeeld aan aandacht voor toetsinstructies voor diegenen die de beoordelingen afnemen. Dergelijke instructies leiden tot een hogere standaardisering van de toetsafname en hebben dus een positieve invloed op de mogelijkheid om de scores te generaliseren.

Dit inzicht brengt ons ertoe ook in ons kwaliteitskader de stappen die voornamelijk te maken hebben met het toetsdesign te scheiden van de elementen die eerder te maken hebben met hoe scores geïnterpreteerd kunnen worden en representatief zijn voor een ruimer domein (toetsdomein, competentiedomein en praktijkdomein), eens de toets is afgenomen.

Wat we meenemen

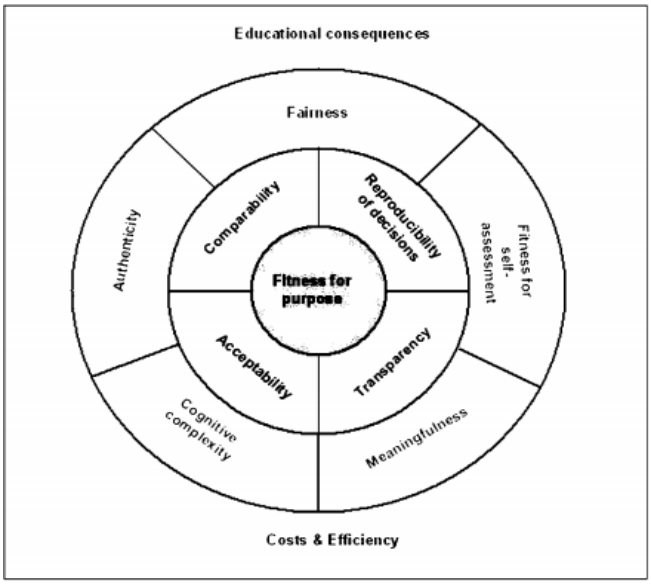
- De aandacht voor ‘domeinbeschrijving’ wordt in onze evaluatiematrix prominent naar voor geschoven.
- Ook ‘toetsafname’ nemen we expliciet als bouwsteen op.
- Om de interpretatieve benadering van Kane ook bruikbaar te maken voor een breder publiek, werken we niet met de volledige argumentatiestructuur, maar kiezen we voor voorwaarden waaraan competentiebeoordelingen op grond van performance assessmenttechieken dienen te voldoen.
- We maken een onderscheid in onze matrix tussen een blok die te maken heeft met de kwaliteit van het toetsdesign en de toets enerzijds; en een blok die te maken heeft met de validiteit van de scores anderzijds.

2.3 Op zoek naar een breder kwaliteitskader

Toetsen en toetsresultaten staan niet los van de context waarin ze plaatsvinden en gebruikt worden. Ook Kane heeft hier aandacht voor in de zin dat, afhankelijk van het doel van de toets, andere argumenten en bewijzen voor die argumenten naar voren kunnen worden geschoven. Ook de laatste stap in de keten van Kane, namelijk het inschatten van de implicaties van de toets, verwijst naar de ruimere context waarbinnen toetsen plaatsgrijpen. Toch wordt in het validiteitskader van Kane in de eerste plaats aandacht besteed aan elementen uit de traditionele opvatting van de notie 'kwaliteit', i.c. de validiteit van (interpretatie en gebruik van) toetsscores, en staan andere kwaliteitselementen minder centraal.

Baartman (2008) pleit in het kader van competentiebeoordelingen voor een kwaliteitskader, dat verder reikt dan de traditionele, psychometrische noties van betrouwbaarheid en validiteit. De edumetrische benadering vormt een alternatief om de specifieke karakteristieken van de beoordelingscultuur (tegenover 'toetscultuur') beter in rekening te brengen (Moss, 1994). Eerder dan positie in nemen voor één van beide benaderingen, zien we de verzoening van beide oogpunten (comprehensieve benadering) als het te volgen pad. In de mate dat de argumentatieve benadering te weinig (expliciet) aandacht besteedt aan zogenaamde alternatieve kwaliteitscriteria, willen we in de ontwikkeling van onze matrix voldoende ruimte inbouwen voor aanvullende kwaliteitscriteria.

Baartman (2008) en collega's (zie ook Baartman, Bastiaens, Kirschner, & van der Vleuten, 2006) vullen dit ruimere kwaliteitskader in op grond van het zogenaamde wiel van competentieassessment (zie Figuur 2). Hiermee bouwen ze voort op het werk van onder andere Linn, Baker en Dunbar (1991).



Figuur 2: Wiel van competentieassessment (bron: Baartman et al., 2006, p. 166).

Niet voor alle contexten zijn alle elementen die in het wiel zijn opgenomen even relevant. Zo kunnen we ons voorstellen dat het kwaliteitscriterium 'meaningfulness' voor een peilingstoets minder belangrijk is dan voor toetsen die in de klas worden ingezet. 'Fairness' en 'transparancy' zijn dan weer wel belangrijk. Centraal in het wiel staat 'fitness for purpose', wat impliceert dat een toets maar kwaliteitsvol kan zijn indien hij geschikt is voor het doel waarvoor hij wordt ontwikkeld. Het betreft een principe dat ook in onze evaluatiematrix verder uitgewerkt zal worden.

Een interessant raamwerk, naast dat van Baartman, wordt ons aangereikt door Newhouse (2011). Teruggrijpend op het werk van Kimbell, Wheeler, Miller en Pollitt (2007) onderscheidt hij zes dimensies die er samen voor zorgen dat een bepaalde toets in een bepaalde context haalbaar (*'*feasible') is. De dimensie 'beheersbaarheid' duidt op de handelbaarheid van de toetsafname, terwijl de 'technische dimensie' heel specifiek verwijst naar technische uitdagingen die voortvloeien uit het inzetten van ICT voor de performance assessment. De 'pedagogische' dimensie heeft te maken met de aanvaarding van de toetsvorm door leerkrachten en leerlingen en met de mate van afstemming op het onderwijs. De twee psychometrische dimensies *(*'functional') gaan in op betrouwbaarheid enerzijds en validiteit anderzijds. Een uitgebreide analyse naar de haalbaarheid gebeurt uiteindelijk op basis van deze verschillende elementen, en is steeds een afweging tussen de verschillende deelcomponenten, rekening houdend met het doel van de toets.

Het afwegen van aspecten komt overigens ook heel duidelijk naar voren in de formule die van der Vleuten en Schuwirth (2005) voorstellen ten aanzien van de kwaliteit en bruikbaarheid van toetsen: *utility of an assessment tool = validity x reliability x acceptability x educational impact x cost-effectiveness*

Zeker bij het opzetten van grootschalige competentietoetsen is dat laatste aspect, namelijk de kosten die bepaalde oplossingen met zich meebrengen, een belangrijk element, dat vaak bepaalt hoe ver men kan gaan (zie o.a. Lane en Stone, 2006). Het is mogelijk om een performance assessment uit te werken die tot zeer betrouwbare scores leidt (bv. studenten moeten vijftig taken uitvoeren en er worden veel beoordelaars ingezet). Of deze performance assessment ook haalbaar is, is maar zeer de vraag. Tijd en middelen van bepaalde toets- en beoordelingsvormen moeten als contextvariabele ook steeds mee in beschouwing worden genomen.

Wat we meenemen

- Samen met Baartman (2008), Newhouse (2011) en Schuwirth en van der Vleuten (2005) plaatsen we het validiteitsraamwerk van Kane in een breder kwaliteitskader, waarin er meer expliciete aandacht is voor o.a. de transparantie en eerlijkheid van de toets en de toetstaken en de impact ervan.
- In het geval van grootschalige performance assessments is het cruciaal dat een ontwikkelde toets ook haalbaar (Newhouse, 2011) en bruikbaar is (Schuwirth & van der Vleuten, 2005). Hier nemen we ook expliciet belangrijke condities als tijd, financiële middelen, infrastructuur (incl. materiaal) en efficiëntie mee in het verhaal.
- De kwaliteit van een toets wordt vastgesteld door de voortdurende afweging van al deze elementen, steeds rekening houdend met het doel van de toets ('fitness for purpose').

3 Methodologie

Teneinde de doelstellingen te realiseren en de beoogde resultaten op te leveren, werden drie onderzoeksstromen opgezet: de systematische literatuurstudie; de selectie en analyse van de praktijkvoorbeelden; en de ontwikkeling van de evaluatiematrix. Bij het uitvoeren van de studie was er steeds sprake van een kruisbestuiving tussen deze onderzoeksstromen.

3.1 Systematische literatuurstudie

De systematisch literatuurreview ontwikkelde zich in verschillende stappen. Voor de opzet van de studie baseerden we ons op Petticrew en Roberts (2008), een standaardwerk inzake het uitvoeren van systematische literatuurstudies in de sociale wetenschappen. Doel was het identificeren van bronnen waarin wordt gerapporteerd over onderzoek naar kwaliteitseisen gesteld aan het toetsen van competenties via performance(-based) assessment. We hielden het blikveld in deze fase bewust breed en namen zowel artikelen mee die rapporteerden over grootschalige toetsen, als artikelen die kleinschaligere initiatieven in kaart brachten.

De eerste stap in de literatuurstudie omvatte het opstellen van een protocol, dat richtinggevend was voor de verdere uitvoering van de review. Het protocol stuurde onder meer het vastleggen van het doel van de literatuurstudie; het afbakenen van de centrale begrippen; het vastpinnen van de fundamentele vraag waarop de review een antwoord dient te geven; en het uittekenen van de aan te wenden zoekstrategie.

Tijdens het selectieproces hielden we nauwgezet de stappen bij die hebben geleid tot het uiteindelijke staal artikelen. De nadruk lag op de kwaliteitscontrole van het selectieproces enerzijds en het coderen en analyseren van de artikelen op 'full-text' anderzijds.

Gedurende het inventarisatieproces wonnen we ook het advies in van buitenlandse experten. De review resulteerde in een staal van 61 artikelen, die finaal gecodeerd werden voor wat betreft het gestelde probleem, de bouwstenen van de evaluatiematrix en - indien beschikbaar - de voorgestelde oplossing.

3.2 Buitenlandse praktijkvoorbeelden

Daarnaast ondernamen we ook stappen voor het verzamelen van praktijkvoorbeelden van bestaande competentiegerichte evaluatiesystemen. In tegenstelling tot de fase van de literatuurreview, focusten we ons hierbij meteen op grootschalige initiatieven gericht op het bewaken van leerlingenprestaties op systeemniveau.

De screening en de selectie van de potentiële praktijkvoorbeelden gebeurde in twee stappen. Op basis van een aantal overzichtswerken (en de informatie die we al sneeuwballend vanuit deze overzichtswerken verzamelden) deden we een eerste grondige screening wat in een, vrij grove, selectie resulteerde. In een tweede stap gebeurde de verdere screening en selectie meer fijnmazig.

De analyse van de praktijkvoorbeelden had enerzijds als doel om de evaluatiematrix verder te verfijnen en op praktische haalbaarheid te toetsen en anderzijds om uitdagingen en eventuele oplossingen bij het grootschalig inzetten van performance assessment voor monitoringtoetsen, in kaart te brengen. Met dit doel voor ogen analyseerden we algemene en technische rapporten en andere relevante documenten. Daarnaast hebben we voor elk van de praktijkvoorbeelden, via skype, twee diepte-interviews opgezet. Terwijl in interviewronde 1 de verschillende bouwblokken van de evaluatiematrix systematisch werden doorlopen, gingen we in de tweede ronde dieper in op een aantal centrale elementen van het toetssysteem. Op basis van de evaluatiematrix werkten we met het oog op de interviews een vragenpool uit (zie bijlage 1 bij het rapport voor de volledige vragenpool), waaruit we voor elke toets afzonderlijk en naargelang de hiaten en vraagtekens, een individuele set vragen samen stelden.

3.3 Genese en validering van de evaluatiematrix

De ontwikkeling van de evaluatiematrix gebeurde in verschillende iteraties. Het denkkader van Kane vormt het startpunt voor de matrixontwikkeling. Op grond van de argumentatieve benadering van validiteit (Kane, 2006, 2013; Kane et al., 1999) ontwikkelden we een eerste versie van de matrix (versie 1.0). Parallel met de ontwikkeling van deze eerste versie, werkten we ook aan de codering van de artikelen uit de literatuurstudie. Bijgevolg vonden de inzichten verkregen uit dat proces en door lectuur van een aantal basiswerken rond 'performance assessment', ook reeds ingang in deze eerste versie. We legden deze eerste versie van de evaluatiematrix (versie 1.0) ter validering voor aan de stuurgroep van het OPBWO-project.

De feedback op versie 1.0 leidde, samen met nieuwe inzichten uit de literatuurstudie, tot een volgende versie van de evalautiematrix (versie 2.0). Deze werd in een tweede valideringsronde voorgelegd aan de vaste expertengroep van het OPBWO-project. Tijdens een vergadering met deze expertengroep in december 2015 werd de evaluatiematrix 2.0 ten gronde besproken. Daarbij zoomden we in op aspecten als kwaliteit, volledigheid, inzichtelijkheid en bruikbaarheid van de voorliggende matrix.

Op grond van feedback van de experten tijdens en na het overleg, inclusief nieuwe inzichten uit de voortschrijdende verkenning van de literatuur, werkten we een volgende versie van de evaluatiematrix (versie 3.0) uit. Deze versie legden we opnieuw ter validering voor aan de leden van de expertengroep. Deze matrixversie vormde bovendien het vertrekpunt voor de interviews en de 'praktische' analyse van de praktijkvoorbeelden. Om de bruikbaarheid, volledigheid en inzichtelijkheid van de evaluatiematrix grondig te toetsen aan de praktijk, werden drie van de praktijkvoorbeelden uitgebreid beschreven volgens de verschillende bouwstenen van de matrix.

De finale evaluatiematrix (versie 4.0) is het resultaat van verschillende iteraties en bundelt heel verscheiden input:

- de wetenschappelijke literatuur: zowel basiswerken, als de systematische literatuurstudie, grijze literatuur en literatuur die voortkwam uit de sneeuwbalmethode;
- input van inhoudsexperten;
- en inzichten vanuit de praktijkvoorbeelden.

Doorheen de verschillende versies van de matrix werd steeds preciezer de vinger gelegd op de essentiële onderdelen van een raamwerk om grootschalige competentietoetsen die gebruik maken van performance assessment op hun kwaliteit te toetsen.

4 Resultaten

In de inleiding van deze beleidssamenvatting schoven we op grond van de vastgelegde doelstellingen, vier concrete resultaten naar voor. Deze resultaten worden hieronder beknopt beschreven.

4.1 Evaluatiematrix 4.0

Bijlage 1 bij deze beleidssamenvatting presenteert de finale evaluatiematrix (versie 4.0) op basis waarvan bestaande en toekomstige toetsprogramma's die competenties beogen te toetsen op grond van performance assessmenttechnieken, op hun kwaliteit beoordeeld kunnen worden. De matrix is opgebouwd op basis van sleutelelementen uit de argumentatieve benaderingsliteratuur, basisinzichten uit het domein Onderwijs & Meten ('educational measurement'), expertise van inhoudsexperten en een praktijktoets op basis van drie praktijkvoorbeelden. De matrix omvat zeven bouwstenen, met daaraan telkens één of meerdere voorwaarden gekoppeld. De evaluatiematrix geeft met andere woorden globaal en per bouwsteen aan, aan welke voorwaarden de peilingstoets dient te voldoen teneinde 1) de bouwstenen zo kwaliteitsvol mogelijk te kunnen neerzetten en 2) op grond van de scores, zo valide mogelijke uitspraken te kunnen doen omtrent het competentiepeil van groepen leerlingen in Vlaanderen.

We volgen de argumentatieve benadering in de zin dat alle verschillende bouwstenen belangrijk zijn in het valideren van (interpretatie en gebruik van) toetsscores. Met het model willen we de afweging beklemtonen die - binnen elk bouwsteen en tussen de bouwstenen - gemaakt moet worden tussen wat de meest kwaliteitsvolle oplossing is, respectievelijk in termen van validiteit en in termen van haalbaarheid (geconcretiseerd in tijd en middelen). Door het expliciteren van de voorwaarden met betrekking tot elke bouwsteen, bouwen we conform het gedachtegoed van Kane een interpretatief- en gebruiksargument op.

Daarnaast volgen we echter ook duidelijk een toetsdesign-insteek; de matrix volgt de logische stappen van het op- en uitzetten van toetsen. De oranje pijl aan de rechterzijde van de matrix geeft deze standaardafwikkeling weer. Het startpunt vormt het peilingsonderzoek als geheel. Pas nadat de bedoeling van de toets duidelijk werd geëxpliciteerd, de beoogde competentie werd gepreciseerd en het toetsdomein werd afgebakend, kristalliseren de voorwaarden zich expliciet en gericht rondom het performance assessment-gedeelte van de toets. De selectie van de meest geschikte toetsvorm hangt immers af van de (dimensies van de) competentie die men wil meten.

De zeven bouwstenen (inclusief voorwaarden) zijn als volgt benoemd:

- doelbepaling *(1 voorwaarde)*
- domeinbeschrijving *(2 voorwaarden)*
- opzet en ontwikkeling *(7 voorwaarden)*
- toetsafname *(1 voorwaarde)*
- scoren *(3 voorwaarden)*
- validiteit *(recapitulatie van 5 voorwaarden)*
- niveaubepaling en rapportering *(2 voorwaarden)*

De bouwsteen 'doelbepaling' staat voor het expliciteren van onder meer het waarom, het wat, het wie en de soort conclusies van de peilingstoets en het peilingsonderzoek. In de domeinbeschrijving draait alles rond het gepreciseerd krijgen van de competentie die men wil meten, inclusief het afbakenen van het toetsdomein vanuit het beoogde competentiedomein. De bouwsteen 'opzet en ontwikkeling' is ruim. Hij omvat, vertrekkend vanuit het toetsdomein, aspecten als taakconstructie (met aandacht voor onder meer authenticiteit), uitwerking van de scoringstool,

toetscompilatie en steekproefopzet (van leerlingen en taken). Na de opzet en ontwikkeling volgt standaard de fase van de toetsafname. De focus van deze bouwsteen ligt op het controleren van variabiliteit in de afnamecondities. Ook de volgende bouwsteen, 'scoren', concentreert zich op potentiële variabiliteit, maar dan met betrekking tot het beoordelen. In deze bouwsteen gaat het voorts ook om het bepalen van afgeleide scores en eventueel ook het equivaleren van de sores (incl. transparantie en doelgerichtheid van de methodes die voor het schalen en equivaleren gehanteerd worden). De bouwsteen 'validiteit', die zowel 'generaliseren' als 'extrapoleren' omvat, heeft een speciaal karakter in die zin dat hij oproept afstand te nemen en bewust te overlopen of wel voldaan werd aan alle voorwaarden om de scores te kunnen generaliseren naar het toetsdomein en vervolgens te extrapoleren naar het beoogde competentiedomein. De voorwaarden in deze bouwsteen concentreren zich met name op het controleren van systematische en toevallige ruis (inclusief de effectiviteit van maatregelen om dit onder controle te houden) en de representativiteit van de taken ten overstaan van het toetsdomein en het beoogde competentiedomein. De laatste bouwsteen in de matrix ten slotte, omvat het rapporteren van de toetsscores en het vastleggen van de prestatiestandaarden waartegen deze scores worden afgezet. Elk van deze bouwstenen wordt geflankeerd door de 'haalbaarheidsvoorwaarde': kwaliteitsvolle oplossingen kunnen enkel worden geïmplementeerd als ze ook haalbaar zijn in termen van tijd en middelen.

4.2 Literatuurstudie

De literatuurreview werd opgezet om een antwoord te bieden op de volgende vraag: *Wat zijn empirisch gefundeerde methoden van performance assessment om competenties te evalueren in het lager, secundair en hoger onderwijs? Aan welke kwaliteitscriteria moet tegemoet gekomen worden? Welke zijn de implicaties voor het beoordelingsbeleid en voor de beoordelingen zelf?*

De systematische analyse van de artikelen tijdens het reviewproces zorgde mee voor input voor het vormgeven van de evaluatiematrix. Een gedetailleerde analyse en codering van elk van de artikelen in fase 1, leverde de basis voor, onder meer, het in kaart brengen van de kwaliteitscriteria voor het beoordelen van competenties via performance assessment. Deze kwaliteitscriteria werden vervolgens geïncorporeerd in (opeenvolgende versies van) de evaluatiematrix in de vorm van verschillende bouwstenen: doelbepaling, domeinbeschrijving, opzet & ontwikkeling, toetsafname, scoren, validiteit: generaliseren, validiteit: extrapoleren en rapportering & niveaubepaling.

Aan het einde van fase 2 categoriseerden we de overblijvende artikelen (n=61) volgens het raamwerk van de evaluatiematrix, om zo een zicht te krijgen op de informatie die we met betrekking tot elk van deze bouwstenen verzameld hadden.

Tabellen 1 en 2 geven een overzicht van de artikelen die bewijsmateriaal aanreiken omtrent kwaliteitscriteria voor het opzetten van performance assessments om competenties te toetsen. In beide tabellen worden de artikelen geclassificeerd volgens de bouwstenen van de evaluatiematrix. Tabel 1 werkt de classificatie verder uit naar verschillende inhoudsdomeinen; tabel 2 naar onderwijsniveau.

	Taal	Weten-schap-pen	Medisch	ICT	Leraren-oplei-ding	Andere	Totaal
Doelbepaling	0	0	0	0	0	0	0
Domeinbeschrijving	1	0	0	0	0	0	1
Opzet en ontwikkeling	4	5	12	0	2	7	30
Toetsafname	1	1	1	0	0	2	5
Scoren	9	4	9	2	3	3	30
Validiteit: generaliseren	10	6	20	2	5	6	49
Validiteit: extrapoleren	6	3	7	1	2	4	23
Rapportering en niveaubepaling	0	0	2	0	0	0	2
Totaal	31	19	51	5	12	22	140

Tabel 1: Classificatie literatuuroogst, naar inhoudsdomein.

	Primair	Secundair	Tertiair	Totaal
Doelbepaling	0	0	0	0
Domeinbeschrijving	1	0	0	1
Opzet en ontwikkeling	4	4	23	31
Toetsafname	1	1	3	5
Scoren	5	8	17	30
Validiteit: generaliseren	5	9	32	46
Validiteit: extrapoleren	3	5	18	26
Rapportering en niveaubepaling	0	0	2	2
Totaal	19	27	95	141

Tabel 2: Classificatie literatuuroogst, naar onderwijsniveau.

De artikelen reikten voornamelijk evidentie aan met betrekking tot de opzet en de ontwikkeling van de toets en het scoren (en inherent ook generaliseren en extrapoleren). Duidelijke leemtes, bouwstenen van de matrix waarover de systematische literatuurstudie weinig tot geen resultaten opleverde, zijn ‘domeinbschrijving’ en ‘rapportering en niveaubepaling’. Daarnaast leidde de systematische analyse van de artikelen ook tot de identificatie van uitdagingen en - eventueel - oplossingen bij grootschalige performance assessments die monitoring op systeemniveau als doelstelling hebben.

4.3 Praktijkvoorbeelden

Zeven internationale praktijkvoorbeelden van grootschalige competentietoetsen die gebruik maken van performance assessment, werden geïnventariseerd en geanalyseerd.

land	naam	domein	jaar
Schotland	Scottish Survey on Literacy and Numeracy *(SSLN)*	geletterdheid (bv. schrijfstuk, groepsdiscussie, …)	2014
Nieuw-Zeeland	National Monitoring Study of Student Achievement *(NMSSA)*	gezondheid & lichamelijke opvoeding	2013
Australië	National Assessment Program Literacy and Numeracy (NAPLAN)	geletterdheid (bv. overtuigend schrijven)	2014
Australië	National Assessment Program *(NAP)* sample assessment	ICT geletterdheid	2014
VS	National Assessment of Educational Progress (NAEP)	wetenschappen	2009
VS	National Assessment of Educational Progress *(NAEP)*	technologie & technische geletterdheid	2014
Nederland	Periodieke Peiling van het Onderwijsniveau (PPON)	schrijfvaardigheid	2009

Tabel 3: Praktijkvoorbeelden van grootschalige toetsprogramma's waarbij gebruik wordt gemaakt van performance assessment: finale pool.

De verzamelde informatie, uit documenten en diepte-interviews, werd beschreven en gesynthetiseerd. Voor drie praktijkvoorbeelden - NAPLAN (Australië), PPON (Nederland) en NMSSA (Nieuw-Zeeland) - gebeurde dit uitgebreid aan de hand van de verschillende voorwaarden uit elk van de bouwstenen van de matrix. De beschrijving van de wijze waarop de verschillende bouwstenen van de matrix in concrete situaties worden ingevuld, welke uitdagingen daarmee verbonden zijn en hoe hiermee wordt omgesprongen, zorgt voor een realistische insteek. Op die manier krijgt de lezer informatie over het toetsprogramma dat bestudeerd wordt. Bovendien zorgen de betreffende praktijkvoorbeelden voor een verdere uitdieping en concretisering van de evidentie die in de verschillende bouwstenen van de matrix in beschouwing genomen kan worden om de kwaliteit van een competentietoets op grond van performance assessment, te beoordelen.

Voor de overige vier praktijkvoorbeelden beperkten we ons tot een summiere weergave van de systemen aan de hand van de bouwstenen van de matrix. Naast de drie geanalyseerde praktijkvoorbeelden en de korte beschrijving van de overige vier praktijkvoorbeelden onder de vorm van samenvattingen, synthetiseerden we een aantal relevante aspecten van de onderzochte praktijkvoorbeelden ook in een overzichtstabel, die werd opgenomen in bijlage 5 bij het rapport. Deze tabel biedt de lezer in een oogopslag de essentiële informatie over de praktijkvoorbeelden.

4.4 Essentiële uitdagingen voor peilingen met een PA-component

De systematische literatuurstudie in combinatie met een doorgedreven analyse van buitenlandse praktijkvoorbeelden, stelden ons in staat de meest cruciale uitdagingen van grootschalige evaluatie van competenties op grond van performance assessment in kaart te brengen. De geraadpleegde bronnen gaven bovendien inzicht in mogelijke werkwijzen en oplossingen die een antwoord bieden op deze uitdagingen. We rapporteren hier samenvattend over deze uitdagingen en oplossingen.

4.4.1 Uitdaging 1: Voldoende taken voorzien

Het feit dat peilingstoetsen vaak een breed domein dienen te bestrijken en het fenomeen van de tussen-taakvariantie leiden tot de vaststelling dat peilingstoetsen een aanzienlijk aantal taken dienen te bevatten om valide en betrouwbaar te zijn. Dit is echter praktisch vaak niet haalbaar in termen van kosten voor de ontwikkeling van de toets en de tijd die leerlingen moeten spenderen aan de toets.

Haalbare en kwaliteitsvolle werkwijzen die in de praktijkvoorbeelden en de wetenschappelijke literatuur aan bod kwamen, zijn:

- de inperking van het competentiedomein naar het toetsdomein via een kwaliteitsvolle domeinbeschrijving;
- het inzetten van matrix-sampling;
- het voorzien van verschillende item-formats in één toets;
- het inzetten van toetsen die meer ingebed zijn in het klasgebeuren.

Deze laatste werkwijze levert mogelijk minder betrouwbare scores op individueel leerlingniveau op, maar onderzoek suggereert dat de technische kwaliteit wel volstaat voor rapportering op groepsniveau.

4.4.2 Uitdaging 2: Standaardisering van toetsafname en scoren

Standaardisering van de toetsafname is een van de manieren om meetfouten tegen te gaan. Hiervoor vallen de meeste toetsprogramma's terug op centraal getrainde toetsassistenten, die dan lokaal worden ingezet. Dit is echter duur en logistiek vaak omslachtig. Vanuit die vaststelling werken sommige systemen met eigen, lokale leerkrachten voor het afnemen van de toets, zij het dan steeds in combinatie met centraal aangestuurde controle en kwaliteitszorg. Digitale systemen kunnen het evenwicht tussen standaardisering en authenticiteit helpen vorm geven. Dit gebeurt door de omgeving waarin leerlingen hun toets afleggen duidelijk af te bakenen en tegelijkertijd door een rijkere en meer authentieke context te bieden. Deze context omvat het gebruik van digitale hulpmiddelen (bv. bronnen op het web) of het bieden van een zekere ruimte aan leerlingen om vrij en flexibel te zoeken naar een oplossing.

Bij het beoordelen van performance assessment is het risico op beoordelaarseffecten groter. Dit is een gevolg van het feit dat menselijke beoordelaars worden ingezet, die een 'judgement' vellen over een prestatie van leerlingen. Empirische studies reiken in dit verband oplossingen aan om toch een degelijke betrouwbaarheid te bekomen, o.a.:

- op grond van een degelijke training gaan beoordelaars consistenter beoordelen;
- verschillende beoordelaars inzetten maakt de meetfout kleiner;
- taken kunnen, onder andere op grond van een 'evidence centered design', zodanig ontworpen worden dat ze consistent gescoord kunnen worden;
- scoringstools kunnen zodanig ontwikkeld worden dat ze dit proces ondersteunen. Zowel analytische als holistische scoringstools kunnen hierbij doelgericht worden ingezet.

Niet al deze oplossingen zijn evenwel haalbaar in termen van middelen en tijd. Het trainen van beoordelaars of de ontwikkeling van eenduidige scoringstools zijn dure en tijdrovende activiteiten; het samenbrengen van beoordelaars en hen aansturen van op afstand brengt vergelijkbare uitdagingen met zich mee. Ook het inzetten van meerdere beoordelaars leidt tot een verhoging van kosten en tijd.

Vanuit deze context bieden zich alternatieve denkrichtingen aan. Paarsgewijze vergelijking lijkt een valide, betrouwbaar en haalbaar alternatief te zijn voor klassiek scoren via rubrics, zeker in combinatie met nieuwe technologische mogelijkheden. Geautomatiseerd scoren doet omwille van een verhoogde efficiëntie zijn intrede, met name bij het beoordelen van schrijfproducten. Niet iedereen is er, vanuit validiteitsoogpunt, echter van overtuigd dat deze laatste werkwijze aan te bevelen is. Net als bij de toetsafname wordt in sommige praktijkvoorbeelden geopteerd om de eigen leerkrachten in te zetten voor het beoordelen. Extra waakzaamheid is dan wel geboden in verband met het optreden van beoordelaarseffecten. Onderzoek lijkt evenwel aan te tonen dat ook hier oplossingen voor kunnen worden geboden, onder andere door systematisch in te zetten op het professionaliseren van leerkrachten.

4.4.3 Uitdaging 3: Vermijden van construct-irrelevante variantie

Construct-irrelevante variantie (CIV) is, zeker bij performance assessments, een belangrijke potentiële bron van systematische meetfouten. We hebben vastgesteld dat de praktijkvoorbeelden hier erg verschillend mee omspringen. Ofwel probeert men deze foutenbron ten allen prijze te vermijden, doorgaans ten koste van de authenticiteit van de toets. Ofwel springt men er iets flexibeler mee om; men probeert de construct-irrelevante variantie in kaart te brengen tijdens de pilootfase van de taken en probeert deze er dan zoveel mogelijk aan te schaven. Ook het inzetten van verschillende meetmethoden ziet men als een mogelijkheid om CIV tegen te gaan.

4.4.4 Uitdaging 4: Het opzetten van taken die recht doen aan de criteriumsituatie

Performance assessments hebben het potentieel om complexe vaardigheden en competenties te meten via authentieke taken. Op die manier kunnen bepaalde constructen meer volledig in kaart worden gebracht. We stellen echter vast dat het voor de geanalyseerde praktijkvoorbeelden niet steeds evident is dit potentieel waar te maken. Complexe taken worden, met het oog op standaardisering, vaak onderverdeeld in verschillende deeltaken, die vervolgens samen iets zeggen over een vaardigheid of competentie. Dit staat tegenover de benadering om complexe vaardigheden en competenties niet zomaar op te vatten als de optelsom van de delen en dus bij voorkeur in hun geheel in kaart te brengen. Het opzetten van authentieke taken brengt sowieso ook een aantal psychometrische uitdagingen met zich mee, onder andere door het feit dat authentieke taken ‘vatbaar’ zijn voor construct-irrelevante variantie.

Opnieuw speelt de afweging tussen standaardiseren en authenticiteit op. Het vastleggen en standaardiseren van aspecten van de toetsprocedure die niet vastgelegd zijn in de criteriumsituatie, leidt tot construct-onderrepresentatie (en dus tot systematische meetfouten). Het loslaten van standaardisering zorgt ervoor dat resultaten minder vergelijkbaar zijn. De oplossing lijkt erop neer te komen zoveel mogelijk trouw te blijven aan de criteriumsituatie, maar terwijl ook een bepaalde graad van standaardisering en controle te behouden. Computergebaseerde toetsen dragen de mogelijkheid in zich dit evenwicht vorm te kunnen geven.

Het meenemen van de criteriumsistuatie in het opzetten van de taak, betekent in principe dat zowel proces als product in kaart worden gebracht. Procescomponenten worden momenteel nog in zeer beperkte mate meegenomen in de praktijkvoorbeelden. Het inzetten van ‘tracking software’ lijkt een beloftevolle piste om bij computergebaseerde toetsen informatie te kunnen verkrijgen over het proces van leerlingen. Deze informatie verkrijgen is echter een middelenintensief proces, dat niet steeds die gegevens opleveren die bijdragen tot het beter in kaart brengen van de competentie van leerlingen.

Het inzetten van performance assessment brengt een meerkost met zich mee. Daarom is het belangrijk erover te waken dat de taken en rubrics die ontwikkeld worden, ook werkelijk de volledige breedte en diepte van het beoogde construct meten. De praktijkvoorbeelden hanteren verschillende methodieken om, tijdens de pilootfase, evidentie hieromtrent te vergaren.

4.4.5 Uitdaging 5: Conform de doelstellingen rapporteren

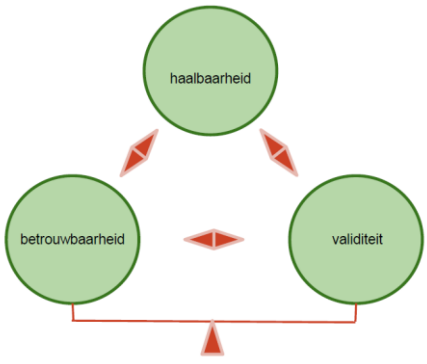
De rapportering vormt het sluitstuk van een peilingstoets. De manier waarop wordt gerapporteerd hangt enerzijds af van de doelstellingen van de toets, maar anderzijds ook van wat haalbaar is.

Bij de analyse van de praktijkvoorbeelden is opgevallen dat IRT zeer frequent wordt gebruikt, zowel om te schalen als met het oog op equivalering. De focus op IRT is in de bestudeerde praktijkvoorbeelden soms zelfs zeer sterk bepalend voor het toetsdesign. In het geval de voorwaarden van IRT geschonden (zullen) worden, beperkt men zich tot het rapporteren van beschrijvende resultaten. Gegeven dat IRT, specifiek voor performance assessment, enkele psychometrische uitdagingen met zich meebrengt, kan men overwegen andere schalings- en equivaleringstechnieken (bijkomend) in te zetten.

Bij het rapporteren is het aangewezen de scores op een betekenisvolle wijze te presenteren. Het vastleggen van prestatiestandaarden is hierin een belangrijke stap, ook bij peilingstoetsen, om te kunnen inschatten welk aandeel van de leerlingenpopulatie de minimumstandaard haalt. Hoewel de nood aan nieuwe, empirisch onderbouwde methoden, gericht op performance assessment, al lang gesignaleerd wordt, valt onder andere uit de literatuurstudie af te leiden dat aan deze oproep slechts beperkt gevolg werd gegeven. Ook bij de analyse van de praktijkvoorbeelden stelden we vast dat innovatieve werkwijzen nog niet zijn uitgewerkt of grondig onderzocht.

4.4.6 Conclusie: generaliseren, extrapoleren, haalbaarheid

Centraal in het opzetten van performance assessment staat het delicate evenwicht tussen de bouwstenen ‘generaliseren’ en ‘extrapoleren’ (de zogenaamde betrouwbaarheid-validiteit paradox); tussen de noodzaak om een accurate, betrouwbare toets op te zetten enerzijds en de ambitie om deze valide en authentiek te maken anderzijds. In het kader van het opzetten van grootschalige toetsen die gebruik maken van performance assessment, stelt zich daarenboven ook de vraag of de gekozen oplossing financieel en logistiek haalbaar is. Niet alleen wat de toetsontwikkeling betreft, maar ook met betrekking tot de afname en de belasting voor leerlingen en leerkrachten. Figuur 3 illustreert de drie grote componenten, te weten betrouwbaarheid (generaliseerbaarheid), validiteit en haalbaarheid, die bij het maken van keuzes in de afweging worden meegenomen.



Figuur 3: Evenwicht tussen componenten van performance assessment.

In de praktijkvoorbeelden zien we dat er zich met betrekking tot de afweging tussen betrouwbaarheid en validiteit twee sporen vormen.

Enerzijds is er een groep systemen waarbij maximale standaardisering het uitgangspunt vormt. We rekenen PPON, NAEP (TEL en Science) en NAPLAN tot deze groepen. Hun primaire doelstelling is om toetsen op te zetten die scores opleveren die vergelijkbaar zijn tussen verschillende toetsafnames op verschillende locaties, tussen verschillende groepen en mogelijk zelfs tussen verschillende afnamejaren. Om die vergelijkbaarheid in de hand te werken, ontwerpen ze taken die duidelijk omlijnd zijn, nemen centraal getrainde toetsassistenten de toets af via strikt uitgewerkte procedures en wordt veel tijd gestopt in het trainen en monitoren van de beoordelaars. In al deze geanalyseerde praktijkvoorbeelden zet men ook authentieke taken in, maar de mate van authenticiteit (van taakinhoud en afname) wordt begrensd door de primaire doelstelling van vergelijkbaarheid.

Anderzijds zijn er praktijkvoorbeelden die validiteit als primair vertrekpunt nemen, zoals bijvoorbeeld NMSSA-Arts en SSLN-Literacy. In deze systemen wordt meer een beroep gedaan op de leerkracht en is de toets van de competentie van leerlingen ingebed in het klasgebeuren. Uitgangspunt is dat gestandaardiseerde toetsen er onvoldoende in slagen de reële competentie van leerlingen aan de oppervlakte te krijgen en door de strikte afgrenzing van de taak ook weinig authentiek zijn. Ook in deze voorbeelden vindt standaardisering plaats, maar wordt deze minder strikt doorgetrokken; enerzijds omdat validiteit de primaire beweegreden is; anderzijds aangezien er gerapporteerd wordt op groepsniveau, wat gevolgen heeft voor de eisen die aan de betrouwbaarheid van de scores voor individuele leerlingen worden gesteld.

Hoewel we voorbeelden hebben gevonden aan beide kanten van het spectrum, is duidelijk dat grootschalige toetsen veelal de kaart van standaardisering trekken, dit met het oog op vergelijkbaarheid van de scores. In deze contexten zijn toetsvormen die meer op authenticiteit gericht zijn, in het verleden uitgetest geweest, maar daarna weer verlaten. Uit de interviews met de praktijkvoorbeelden blijkt dat sommige systemen in het verleden hebben geëxperimenteerd met meer authentieke toetsvormen (bv. portfolio's). Dit leverde echter te veel oncontroleerbare variantie op en bleek niet werkbaar, te meer omdat de doelstelling soms ook was om resultaten over verschillende afnamejaren heen te vergelijken. Men wees expliciet op de grote uitdagingen verbonden aan het inzetten van lokale leerkrachten, onder andere ook in het perspectief van het realiseren van (de deelname van) de vooropgestelde steekproef van leerlingen. Bovendien merkte men ook op dat het net één van de uitgangspunten is om een systeem op te zetten dat zo weinig mogelijk extra inspanningen vergt van leerkrachten en leerlingen en dat systemen die de kaart van constructvaliditeit trekken daar net niet in slagen.

Tegen de achtergrond van deze kritiek stelt zich de vraag of de hoge betrouwbaarheidscoëfficiënten die nagestreefd worden bij toetsen met een hoge inzet voor de individuele leerling ('high stakes'), ook een voorwaarde zijn bij rapportage op geaggregeerd niveau, met name in het geval van grootschalige toetsprogramma's die kwaliteitszorg op systeemniveau beogen. Volgens een recente publicatie van de National Research Council (2014) wordt meer en meer erkend dat "reliability statistics for individual-level scores and decisions are different from those for higher levels of aggregations" (p. 189). De publicatie verwijst naar werk van Hill en DePascale (2003), dat aantoont dat betrouwbaarheidsniveaus die een bron van zorg kunnen zijn bij rapportering op individueel niveau, nog steeds vaststellingen op hogere niveaus, zoals het schoolniveau, kunnen ondersteunen. In de casus Nieuw-Zeeland werd tijdens de interviews omtrent NMSSA heel duidelijk aangegeven dat ze een grotere marge hebben, omdat ze enkel op groepsniveau rapporteren. In de 'Standards' (AERA, APA, & NCME, 2014) lezen we op dit vlak dat het vasthouden aan een gepaste mate van standaardisering van de afnameprocedures vooral haar belang heeft bij toetsen waar voor de leerling(en) in kwestie, veel op het spel staat ('high stakes').

Duidelijk is enerzijds dat men in andere systemen keuzes maakt op basis van de insteek die volgens hen primeert: standaardisering of validiteit. De gemaakte keuzes hebben aan de ene kant te maken met de doelstellingen van de

toets; anderzijds ook met de heersende cultuur. Zo hoorden we tijdens de interviews dat het niet toevallig is dat net Schotland en Nieuw-Zeeland met alternatieve werkwijzen experimenteren. In Nieuw-Zeeland zet men sinds enkele decennia in op professionalisering van leerkrachten met het oog op formatieve evaluatie. Binnen een dergelijke context is er meer vertrouwen in leerkrachten om prestaties van leerlingen in te schatten, te meer omdat leerkrachten in Nieuw-Zeeland dat in andere toetsprogramma's ook reeds doen. In Schotland zien we een gelijkaardige evolutie en zijn leerkrachten, door het jarenlang werken met goed gedefinieerde prestatieniveaus, veel empirischer bezig met deze niveaus. Tegen deze achtergrond, zo stelden ook de mensen bij PPON, zijn ze beter in staat een overwogen selectie te maken van schrijfstukken.

Uit de publicatie van de National Research Council (2014) leiden wij af dat ook de Verenigde Staten de deur open houden voor nieuwe ontwikkelingen. In het interview met betrekking tot NAEP-TEL werd aangestipt dat ideale werkwijzen meer in de richting gaan van combinaties van beide 'uitersten', waarbij delen van een toets via strikt gestandaardiseerde werkwijzen verlopen, en andere onderdelen meer ruimte laten voor alternatieven. Ook bij de uitwerking van de uitdagingen en vaststellingen in dit hoofdstuk, kwamen gecombineerde aanpakken al naar voren als mogelijke oplossing voor welbepaalde problemen, zoals bijvoorbeeld het inzetten van een mix van item- en taaktypes of het combineren van toetsen op vraag, met toetsen die ingebed zijn in de klaspraktijk.

Uit de analyses komt duidelijk naar voren dat het opzetten van een toets steeds een zoektocht is naar de beste en efficiëntste manier om de beste en rijkste data te verzamelen, die aan antwoord bieden op de gestelde vragen, conform de doelstelling van de toets. Er is sprake van een afweging; een keuze voor de ene component betekent dat ook opofferingen gemaakt moeten worden met betrekking tot een andere component

5 Beleidsaanbevelingen

Uit het onderzoek kunnen lessen getrokken worden voor het actuele en het toekomstige beleid. In dit laatste deel van het rapport formuleren we bijgevolg elf aanbevelingen die naar aanleiding van dit onderzoek aan de oppervlakte kwamen. Deze set aanbevelingen wordt opgehangen aan twee perspectieven van waaruit de aanbevelingen kunnen worden beschouwd: enerzijds de eindtermen die als minimale doelstellingen het startpunt vormen van het peilingsonderzoek in Vlaanderen; anderzijds het (toekomstige) competentiegerichte peilingsonderzoek zelf.

Elke aanbeveling staat niet alleen op zichzelf, maar is ook steeds gerelateerd aan de overige. Vanuit een korte situatieschets komen we telkens tot een concrete aanbeveling; de ruimere context werd reeds uitgebreid besproken in deel 3 van het rapport. We kunnen in dit verband zelfs stellen dat er reeds op basis van het resultatenluik duidelijke implicaties te destilleren vallen (zie o.m. deel 3, 4.).

5.1 Formulering van de eindtermen

In lijn met het principe van 'constructive alignment' zijn voor de ontwikkeling van toetssystemen die in staat zijn om op een kwaliteitsvolle manier competenties te meten, leerdoelen nodig die competentiegericht zijn geformuleerd en die aansluiten bij competentiegerichte eindtermen. Voor het peilingsonderzoek in Vlaanderen vormen de eindtermen het vertrekpunt. Op dit moment zijn de meeste eindtermen echter vooral op een kennisgerichte leest geschoeid en matig doordrongen van het gedachtegoed inzake competentiegerichtheid, alhoewel er ook al competentiegerichte eindtermen zijn geformuleerd. Daarnaast stellen we vast dat de competentiegerichte benadering ook in verschillende mate ingang heeft gevonden in de onderwijspraktijk zelf, al naargelang van het onderwijsniveau, de onderwijsvorm en de studierichting.

Deze situatie vormt het aangrijpingspunt voor het pleidooi om – indien de maatschappelijke vraag aanwezig is om zicht te krijgen op bereikte competenties bij leerlingen in verschillende stadia van hun studieloopbaan – te (blijven) werken aan een competentiegerichte invulling van de eindtermen en vooropstellen van leerdoelen in het onderwijs. Indien het de bedoeling is om de peilingstoetsen meer competentiegericht te maken, wijst voorliggend onderzoek er op dat de huidige invulling van de eindtermen momenteel geen adequaat vertrekpunt vormt voor de fase van de doelbepaling. Gezien het belang van de doelbepaling impliceert dit dat de randvoorwaarden niet vervuld zijn voor het opzetten van kwaliteitsvolle toetsen. Hier kan in de huidige context alleen aan tegemoet gekomen worden door een afzonderlijke inspanning van toetsontwikkelaars om de eindtermen, in opdracht van de overheid en in samenspraak met het onderwijsveld, te herformuleren naar competentiegerichte doelen, als vertrekpunt voor de uitwerking van het toetsprogramma.

Aanbeveling 1: Als de overheid ervoor opteert om peilingen van *competenties* op te zetten, dan zijn eindtermen geformuleerd onder de vorm van competenties noodzakelijk.

De eindtermen vormen enerzijds de minimumdoelen die de overheid noodzakelijk en bereikbaar acht voor een bepaalde leerlingenpopulatie. Anderzijds zijn zij het vertrekpunt van waaruit toetsprogramma's en toetsen worden opgezet. In het licht van hun rol als aanknopingspunt voor toetsontwikkeling is het aangewezen om bij de ontwikkeling van de eindtermen ook het perspectief van het kwaliteitsvol en haalbaar toetsen mee te nemen, met het oog op de discussie bij het vertalen van deze doelen naar toetsbare eenheden. Daarenboven is het aangewezen om bij de formulering van elke individuele eindterm tevens de ambitie te formuleren of deze al dan niet grootschalig toetsbaar moet zijn. Immers, het zou onverstandig zijn om 'grootschalige toetsbaarheid' als conditio sine qua non

naar voor te schuiven bij het formuleren van (competentiegerichte) eindtermen. Het is aangewezen om daarom, naast andere experts, ook toetsontwikkelaars en/of toetsexperts uit bv. het Steunpunt Toetsontwikkeling en Peilingen te betrekken of te raadplegen bij het uitwerken van voorstellen tot competentiegerichte eindtermen, zodanig dat vanuit deze expertise de grenzen aan het grootschalig toetsen van deze eindtermen geduid kunnen worden. Deze informatie is verrijkend voor de discussie en de beslissingen van de overheid bij het herformuleren van eindtermen.

Aanbeveling 2: Bij de formulering van (competentiegerichte) eindtermen verdient het de aanbeveling om ook toetsdeskundigen te betrekken.

Het consequent doortrekken van het principe van 'constructive alignment' houdt in dat de overheid ook oog dient te hebben voor de gevolgen van een herformuleren van de eindtermen in de onderwijspraktijk, meer bepaald de consequenties ten aanzien van de didactische aanpak en evaluatiepraktijk van leerkrachten in de klas. Een evolutie richting competentiegerichte eindtermen impliceert immers dat leerkrachten meer dan voorheen de kaart zullen (moeten) trekken van competentiegericht onderwijs én competentiegericht beoordelen van leerlingen. Het beleid zal in dat geval lacunes en behoeften op dat vlak in kaart moeten brengen en verdere professionalisering van leerkrachten stimuleren. Dit onderzoek - en de bijhorende praktijkbrochure - kan alvast input leveren voor het uitwerken van een professionaliseringsaanbod.

Uit dit onderzoek blijkt bovendien de aanwezigheid van een lerarenkorps dat een professionaliseringstraject heeft doorlopen gericht op het beoordelen van leerlingen in het licht van (nationale) onderwijsdoelstellingen en standaarden, de deur opent naar alternatieve opzetten van grootschalige competentietoetsen (zie bv. NMSSA-Arts in Nieuw-Zeeland).

Aanbeveling 3: De keuze voor competentiegerichte eindtermen impliceert een actief beleid rond het professionaliseren van leerkrachten inzake competentiegericht onderwijs met in het bijzonder aandacht voor het competentiegericht toetsen.

5.2 Competentiegericht peilingsonderzoek in Vlaanderen

Een evolutie naar meer competentiegericht onderwijs heeft tot gevolg dat competenties mee in het vizier komen van peilingsonderzoek. Performance assessment blijkt een krachtige manier om deze competenties te toetsen, onder andere omwille van het potentieel om leerlingen complexe taken te laten uitvoeren in een zo levensecht mogelijke context.

Uit ons onderzoek blijkt dat het inzetten van performance assessment bij het grootschalig toetsen van competenties weloverwogen dient te gebeuren. De evaluatiematrix die we in het kader van dit onderzoek ontwikkelden vestigt duidelijk de aandacht op de vraag rond de te hanteren toetsvorm. Pas nadat de bedoeling van de toets duidelijk werd geëxpliciteerd, de beoogde competentie is verfijnd en het toetsdomein is afgebakend, kan een weloverwogen keuze gemaakt worden over de te gebruiken toetsvormen.

Het is zaak goed na te denken in welke mate en/of met betrekking tot welke dimensies van de beoogde competentie performance assessment kan worden ingezet. De keuze om performance assessment in te zetten impliceert met

andere woorden niet dat voor korte invulvragen en/of meerkeuzevragen geen ruimte meer is. Elke toetsvorm heeft duidelijke voor- en nadelen en deze dienen te worden afgewogen tegen het doel van de toets dat eerder werd vastgelegd. Performance assessment dient effectief een meerwaarde op te leveren ten opzichte van standaard toetsvormen, zeker in het licht van de meerkost (bv. in termen van tijd, middelen en inzet van beoordelaars) die daaraan verbonden is.

In de bestudeerde praktijkvoorbeelden zien we naast toetssystemen die louter uit performance assessment bestaan, ook verschillende voorbeelden waarin competenties of complexe vaardigheden getoetst worden aan de hand van een mix van toetsvormen (bv. meerkeuzevragen naast PA-taken). Deze werkwijze heeft zowel vanuit kwaliteitsoogpunt als naar haalbaarheid toe, positieve effecten. Het gebruiken van verschillende itemformats levert naar validiteit van scores toe, voordelen op. Elke specifieke toetsvorm brengt immers welbepaalde meetfouten met zich mee en door toetsvormen te combineren, middelt men deze specifieke methode-effecten uit en wordt construct-irrelevante variantie voor een stuk onder controle gehouden. Wat haalbaarheid betreft, biedt het combineren van toetsvormen de mogelijkheid om brede constructen in een verantwoorde tijdspanne te toetsen.

Aanbeveling 4: De beslissing om performance assessment in te zetten bij grootschalige competentietoetsen gericht op monitoring op systeemniveau - al dan niet in combinatie met andere toetsvormen - moet doelgericht zijn.

Bovenstaande aanbeveling benadrukt opnieuw de logica van de doelstellingen als vertrekpunt voor de verdere ontwikkeling van de toets. De vraag of een toets kwaliteitsvol is ingevuld, kan enkel beantwoord worden door te kijken of beslissingen in het ontwikkelproces in lijn liggen met de doelstellingen die eerder in de fase van de doelbepaling geëxpliciteerd werden. Een kwaliteitsvolle toets ontwikkelen begint met andere woorden met een gestructureerde doelbepaling, die uit verschillende deelcomponenten bestaat: waarom gaan we een toets(programma) opzetten, wat willen we op grond daarvan meten (en bij wie), en welke conclusies willen we daaruit kunnen trekken (onder welke vorm)? Keuzes die men met betrekking tot elk van deze deelcomponenten maakt, beïnvloeden elkaar wederzijds. Bovendien hebben ze ook gevolgen voor wat betreft de verdere ontwikkeling van de toets.

De analyse van internationale praktijkvoorbeelden die we voor dit onderzoek uitvoerden toont aan dat het helder krijgen en beantwoorden van bovenstaande vragen vaak een taak is die door de overheid als opdrachtgever zelf wordt uitgevoerd. Hierbij wordt een breed draagvlak gezocht door uiteenlopende actoren bij de discussies te betrekken, zodat er een voldoende breed beleidsdraagvlak ontstaat voor de doelstellingen die vastgelegd worden. De overheid schrijft daarbij pas een aanbesteding uit voor het ontwikkelen en uitvoeren van peilingsonderzoek nadat ze alle beslissingen in de doelbepaling vastlegde.

De rol van de opdrachtgever bij het uitwerken van een duidelijk afgelijnd doel van de toets heeft organisatorische implicaties. De bestudeerde praktijkvoorbeelden tonen aan dat een degelijke omkadering een vereiste is indien men de ambitie van een duidelijke doelbepaling door de opdrachtgever wil waarmaken.

Aanbeveling 5: Zorg in de aanbesteding van grootschalige competentietoetsen met een PA-component, gericht op monitoring op systeemniveau voor een heldere en volledige doelbepaling en reserveer als overheid voldoende tijd en middelen om deze doelbepaling helder te krijgen.

Uit de analyse van de praktijkvoorbeelden leren we dat toetsprogramma's en toetsen multipele doelen kunnen dienen. Verschillende overwegingen kunnen aan de basis daarvan liggen. Grootschalige toetsen vergen een aanzienlijke inspanning in termen van tijd en middelen, wat leidt tot de logische overweging of met één toets niet verschillende vragen beantwoord kunnen worden. Een andere reden is dat grootschalige toetsen enkel afgenomen kunnen worden met medewerking van scholen en leerlingen en dat daarom, in het kader van kwaliteitsbewaking op leerling- en schoolniveau, ook nagedacht kan worden over nuttige informatie die aan scholen aangeleverd kan worden.

Ons onderzoek toont echter aan dat waakzaamheid geboden is bij toetsen die een hybride doelbepaling hebben, net omdat aan deze uiteenlopende doelstellingen andere kwaliteitsvereisten voor het opstellen van de toets verbonden zijn. Vanuit het perspectief om ook in het onderwijsveld een draagvlak voor een toetssysteem te creëren, is het bijvoorbeeld perfect te verdedigen dat scholen die deelnemen aan grootschalige toetsen met het oog op kwaliteitsmonitoring op systeemniveau, ook informatie krijgen over de prestaties van de eigen school en zelfs van individuele leerlingen. Het risico bestaat dan echter dat het toetsprogramma noch de toets initieel opgezet werden met deze bijkomende doelstellingen voor ogen en dat de resultaten niet voldoende betrouwbaar zijn op het niveau van de school of de individuele leerling. Hoewel de opdrachtgever hiermee kan omgaan door bijvoorbeeld in de rapporten voor individuele scholen duidelijk aan te geven welke de beperkingen van de resultaten zijn, leren buitenlandse voorbeelden ons dat deze resultaten soms een eigen leven kunnen gaan leiden en dat toetssystemen die in principe zijn opgezet in een 'low stakes'-context, toch als 'high stakes' beschouwd worden. Gevolg: de 'nieuwe' interpretatie van de resultaten (in dit geval schoolniveau i.p.v. systeemniveau) is niet meer (geheel) valide.

Aanbeveling 6: Wees bij grootschalige competentietoetsen met een PA-component, gericht op monitoring op systeemniveau waakzaam in het geval van 'hybride doelstellingen' en overdenk de gevolgen hiervan voor toetsopzet, rapportering én het gebruik van resultaten.

Grootschalige competentietoetsen, opgezet vanuit het oogpunt de kwaliteit van het onderwijs op systeemniveau te meten moeten betrouwbare en valide resultaten opleveren, teneinde het beleid gefundeerd te kunnen informeren. In dit onderzoek werd nagegaan op basis van welke bouwstenen en voorwaarden, grootschalige competentietoetsen met een PA-component, kwaliteitsvol uitgewerkt kunnen worden.
De evaluatiematrix die het resultaat van dit onderzoek is, omvat zeven bouwstenen. Toekomstige grootschalige competentietoetsen kunnen afgetoetst worden aan de kwaliteitsvoorwaarden die in elke afzonderlijke bouwsteen van de matrix geëxpliciteerd worden.

Aanbeveling 7: **Maak gebruik van de bouwstenen en voorwaarden geïdentificeerd in de evaluatiematrix om te bepalen of grootschalige competentietoetsen met een PA-component, gericht op monitoring op systeemniveau, kwaliteitsvol zijn.**

Door in te spelen op de voorwaarden uit de evaluatiematrix kan in principe een kwaliteitsvolle toets worden uitgewerkt. Deze studie toont echter aan dat voor het realiseren van een kwaliteitsvolle toets, keuzes gemaakt moeten worden. De matrix vormt het ideaalplaatje; de uiteindelijke toets is een doordruk van dat plaatje in de werkelijkheid, waarbij de voorwaarden met betrekking tot elk van de bouwstenen met elkaar afgewogen worden, rekening houdend met het doel van de toets. Op dat vlak kunnen spanningen optreden tussen wat wenselijk is en feitelijk haalbaar. Zo is het bijvoorbeeld niet realistisch te verwachten dat grootschalige competentietoetsen, die - in een ideaalscenario - geheel betrouwbare en valide scores opleveren, ook nog eens eenvoudig haalbaar blijken te

zijn in termen van vereiste tijd en middelen. De drie centrale componenten die, bij het maken van keuzes inzake opzet en uitvoering van grootschalige competentietoetsen op basis van performance assessment, met elkaar afgewogen moeten worden zijn: generaliseerbaarheid, extrapoleerbaarheid en haalbaarheid (in termen van tijd en middelen). Het delicate evenwicht tussen generaliseren en extrapoleren wordt ingegeven door de noodzaak om een accurate, betrouwbare toets op te zetten enerzijds en de ambitie om deze zo authentiek en valide mogelijk te maken anderzijds. De initiatieven met het oog op de generaliseerbaarheid van de scores, zoals standaardisering van de toets en het voorzien van grote steekproeven (o.m. van taken, beoordelaars, afnamemomenten, …), blijken in realiteit moeilijk te combineren met maatregelen die de extrapoleerbaarheid van de scores beogen, zoals het uitwerken van authentieke taken die recht doen aan de criteriumsituatie. In het kader van het opzetten van grootschalige toetsen die gebruik maken van performance assessment, stelt zich daarenboven ook de vraag of de toets financieel en logistiek haalbaar is.

Aanbeveling 8: Maak bij het realiseren van grootschalige competentietoetsen met een PA-component, gericht op monitoring op systeemniveau, een weloverwogen afweging tussen generaliseerbaarheid van scores, extrapoleerbaarheid van scores en haalbaarheid (in termen van tijd en middelen).

De afweging tussen de mogelijkheid tot generaliseren enerzijds en tot extrapoleren anderzijds, houdt in zich dat het standaardiseren van de toets deels ten koste gaat van de validiteit van de toets, en ook andersom. De vaststelling in deze studie is dat, met betrekking tot deze afweging, in de meeste bestudeerde praktijkvoorbeelden de kaart wordt getrokken van doorgedreven standaardisering, ten koste van de validiteit waarop PA's in principe aanspraak kunnen maken. Dit doet de vraag rijzen welke mate van standaardisering in feite wenselijk en noodzakelijk is.

Deze studie toont aan dat er ook een alternatieve piste mogelijk is, waarbij de doorgedreven standaardisering van de afname en het scoren van de toets voor een stuk wordt losgelaten door lokale leerkrachten in te zetten om de toets af te nemen en zelfs te scoren. Deze werkwijze heeft enerzijds voordelen op het vlak van validiteit, onder andere in de zin dat leerkrachten beter kunnen inschatten wat het reële competentieniveau van hun leerlingen is. Anderzijds zijn er logistieke en financiële voordelen, bijvoorbeeld omdat het werken met centraal getrainde toetsassistenten, onder meer ook omwille van de logistiek, duur is.

Het grootste nadeel van deze werkwijze is dat de scores mogelijk minder generaliseerbaar (betrouwbaar) zijn. Meer en meer echter, wordt erkend dat er een verschil bestaat tussen betrouwbaarheidsstatistieken voor scores en beslissingen op individueel niveau, vergeleken met deze op hogere aggregatieniveaus. Onderzoek toont immers aan dat betrouwbaarheidsniveaus die een bron van zorg kunnen zijn bij rapportering op individueel niveau, nog steeds vaststellingen op hogere niveaus, kunnen ondersteunen. Bovendien weten we ook dat het vasthouden aan een gepaste mate van standaardisering van de afnameprocedures vooral haar belang heeft bij toetsen waar voor de leerling(en) in kwestie, veel op het spel staat ('high stakes').

We stellen op grond van de praktijkvoorbeelden vast dat aan het slagen van deze alternatieve werkwijze een aantal voorwaarden zijn gekoppeld. Ten eerste dient de inzet van lokale leerkrachten voor het afnemen van de toets gecombineerd te worden met centraal aangestuurde controle en kwaliteitszorg. Een tweede voorwaarde verbonden aan het inzetten van lokale leerkrachten is dat er (verder) werk wordt gemaakt van de professionalisering van leerkrachten inzake toetsen en evalueren.

Ten slotte blijkt uit ons onderzoek dat deze werkwijze zowel voor- als tegenstanders heeft. Voorstanders geven aan meer belang te hechten aan het uitwerken van valide toetsen, terwijl tegenstanders meer de nood aan betrouwbare resultaten benadrukken. Tegen die achtergrond is het belangrijk dat de opdrachtgever klaarheid schept over waar voor een welbepaalde toets de nadruk dient te liggen. Het maken van deze keuze kan deel uitmaken van de doelbepaling.

Aanbeveling 9: Sta bij het opzetten van grootschalige competentietoetsen met een PA-component, gericht op monitoring op systeemniveau, open voor andere opties dan strikt gestandaardiseerde toetssystemen.

Dat ook gekeken wordt naar alternatieve manieren om vorm te geven aan kwaliteitsvolle, grootschalige competentietoetsen, neemt niet weg dat deze uitgewerkt dienen te worden vanuit een streven naar een ideaal evenwicht tussen de mogelijkheid tot generaliseren enerzijds en extrapoleren anderzijds. In dit onderzoek werden met het oog op het bereiken van dit evenwicht een aantal uitdagingen geïdentificeerd, waarvoor een oplossing dient te worden gezocht.

De uitdaging met betrekking tot de mogelijkheid om scores te generaliseren, heeft te maken met het voorzien van voldoende taken. De constructen die via peilingstoetsen gemeten worden, bestrijken vaak een breed domein. Gecombineerd met de problematiek van de tussen-takenvariantie zorgt dit ervoor dat peilingstoetsen een aanzienlijk aantal taken dienen te bevatten om betrouwbare en valide scores op te leveren. Dit is echter praktisch vaak niet haalbaar in termen van kosten verbonden aan de ontwikkeling van de toets en de tijd die leerlingen moeten spenderen aan de toets. Een oplossing die in de geanalyseerde praktijkvoorbeelden en in de literatuur veel gebruikt wordt, is matrix sampling. Bij deze techniek worden steekproeven van taken uit de totale takenpool afgenomen bij steekproeven leerlingen.

Aanbeveling 10: **Overweeg bij grootschalige competentietoetsen met een PA-component, gericht op monitoring op systeemniveau, om matrix sampling te gebruiken.**

In de uitwerking van deze studie hebben we vastgesteld dat zowel het onderzoek naar, als de praktijk van het grootschalig toetsen van competenties op basis van performance assessment, snel evolueert. Er bieden zich beloftevolle pistes aan, die een antwoord bieden voor een aantal essentiële uitdagingen waar deze toetsprogramma's en toetsen mee te kampen hebben. Een aantal van deze pistes werden in dit onderzoek geïdentificeerd.

Paarsgewijze vergelijking lijkt een valide, betrouwbaar en haalbaar alternatief te zijn voor scoren van PA-taken via specifieke scoringstools, zeker in combinatie met nieuwe technologische mogelijkheden. Daarnaast doet geautomatiseerd scoren omwille van het efficiëntievoordeel zijn intrede, met name bij het beoordelen van schrijfproducten. Niet iedereen is er, vanuit validiteitsoogpunt, echter van overtuigd dat deze laatste werkwijze aan te bevelen is. Ook het inzetten van lokale leerkrachten voor toetsafname en scoren, is een piste die volop wordt verkend, om oplossingen te vinden in termen van validiteit en haalbaarheid.

Uit het onderzoek kwam bovendien naar voren dat digitale systemen het evenwicht tussen standaardisering en authenticiteit mee kunnen helpen vorm geven. Dit gebeurt door de omgeving waarin leerlingen hun toets afleggen duidelijk af te bakenen en tegelijkertijd door een rijkere en meer authentieke context te bieden. Deze context omvat

het gebruik van digitale hulpmiddelen (bv. bronnen op het web) of het bieden van een zekere ruimte aan leerlingen om vrij en flexibel te zoeken naar een oplossing.

Net omdat onderzoek niet stil staat en nieuwe inzichten uit empirisch onderzoek in de praktijk uitgetest worden, is de verwachting dat in de komende jaren nieuwe evidentie zal opduiken met betrekking tot de diverse bouwstenen van de evaluatiematrix en de uitdagingen verbonden aan een kwaliteitsvolle invulling ervan. Het is belangrijk om hierover als overheid de vinger aan de pols te houden. Tijdens de uitvoering van deze studie viel het ook op dat vele van de buitenlandse praktijkvoorbeelden bereid zijn om inzichten en ideeën te delen en nieuwe richtingen momenteel worden verkend en in de toekomst zullen worden geëvalueerd. Het vormen van een internationaal netwerk voor kennisdeling, lijkt bijgevolg een van de mogelijkheden om op de hoogte te blijven van recente ontwikkelingen.

Aanbeveling 11: Blijf oog hebben voor nieuwe ontwikkelingen in onderzoek naar en de praktijk van grootschalige competentietoetsen met een PA-component, gericht op monitoring op systeemniveau.

Bibliografie

AERA - American Educational Research Association, APA - American Psychological Association, & NCME - National Council on Measurement in Education (2014). *Standards for Educational and Psychological Testing.* Washington, DC: American Educational Research Association.

Baartman, L. (2008). *'Assessing the assessment': Development and use of quality criteria for Competence Assessment Programmes.* (Unpublished doctoral dissertation). Faculty of Education, University of Utrecht, the Netherlands.

Baartman, L., Bastiaens, T., Kirschner, P., & van der Vleuten, C. (2006). The wheel of competency assessment: Presenting quality criteria for competency assessment programs. *Studies in Educational Evaluation, 32*(2), 153-170.

Brennan, R. (2006). *Educational measurement.* Westport, CT: American Council on Education: Praeger.

Chapelle, C. (2012). Validity argument for language assessment: The framework is simple. *Language Testing, 29*(1), 19.

Chapelle, C., Enright, M., & Jamieson, J. (2010). Does an Argument-Based Approach to Validity Make a Difference? *Educational measurement: issues and practice, 29*(1), 3-13.

Crisp, V., & Shaw, S. (2011). Applying methods to evaluate construct validity in the context of A level assessment. *Educational Studies, 38*(2), 209-222.

Crooks, T., Kane, M., & Cohen, A. (1996). Threads to the valid use of assessment. *Assessment in Education, 3*(3), 265-285.

Curcin, M., Boyle, A., May, T., & Raman, Z. (2014). A validation framework for work-based observational assessment in vocational qualifications. Coventry: Office of Qualifications and Examinations Regulation.

Enright, M., & Quinlan, T. (2010). Complementing human judgment of essays written by English language learners with e-rater® scoring. *Language Testing*, *27*(3), 317-334.

Evers, A., Lucassen, W., Meijer, R., & Sijtsma, K. (2009). *COTAN beoordelingssysteem voor de kwaliteit van tests (geheel herziene versie)*. Amsterdam: NIP/COTAN.

Hill, R. K., & DePascale, C. A. (2003). Reliability of No Child Left Behind Accountability Designs. *Educational Measurement: Issues and Practices, 22*(3), 12-20.

Kane, M. (2004). Certification Testing as an Illustration of Argument-Based Validation. *Measurement: Interdisciplinary Research and Perspectives, 2*(3), 135-170.

Kane, M. (Ed.). (2006). *Validation* (4 ed.). Westport: Praeger Publishers.

Kane, M. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement, 50*(1), 1-73.

Kane, M., Crooks, T., & Cohen, A. (1999). Validating measures of performance. *Educational measurement: issues and practice, 18*(2), 5-17.

Kimbell, R., Wheeler, T., Miller, A., & Pollitt, A. (2007). *E-scape: e-solutions for creative assessment in portfolio environments.* London: Technology Education Research Unit, Goldsmiths College.

Lane, S., & Stone, C. (2006). Performance Assessment In R. Brennan (Ed.), *Educational measurement* (4 ed., pp. 387-432). Westport: American Council on Education/Praeger.

Linn, R., Baker, E., & Dunbar, S. (1991). Complex, performance-based assessment: Expectations and validation criteria. *Educational researcher, 20*(8), 15-21.

Llosa, L. (2008). Building and Supporting a Validity Argument for a Standards-Based Classroom Assessment of English Proficiency Based on Teacher Judgments. *Educational measurement: issues and practice, 27*(3), 32-42.

Moss, P. (1994). Can there be validity without reliability? *Educational researcher, 23*(2), 5-12.

National Research Council (2014). *Developing Assessments for the Next Generation Science Standards. Committee on Developing Assessments of Science Proficiency in K-12.* Washington, DC: The National Academies Press.

Newhouse, C. P. (2011). Using IT to assess IT: Towards greater authenticity in summative performance assessment. *Computers & Education, 56*(2), 388-402.

Petticrew, M. & Roberts, H. (2008). *Systematic reviews in the Social Sciences: A Practical Guide*. Malden, VS: Blackwell.

Schuwirth, L., & van der Vleuten, C. (2012). Programmatic assessment and Kane's validity perspective. *Medical education, 46*(1), 38-48.

Shaw, S., Crisp, V., & Johnson, N. (2011). A framework for evidencing assessment validity in large-scale, high-stakes international examinations. *Assessment in Education: Principles, Policy & Practice, 19*(2), 159-176.

Toulmin. (1958). *The uses of argument.* Cambridge: Cambridge University Press.

van der Vleuten, C., & Schuwirth, L. (2005). Assessing professional competence: from methods to programmes. *Medical education, 39*(3), 309-317.

Wools, S. (2012) Towards a comprehensive Evaluation System for the Quality of Tests and Assessments *Psychometrics in Practice at RCEC* (pp. 95-106). Enschede: RCEC.

Wools, S. (2015). *All About Validity - An evaluation system for the quality of educational assessment.* Enschede. (Unpublished doctoral dissertation). University of Twente, the Netherlands.

Wools, S., Sanders, P., & Roelofs, E. (2007). *Beoordelingsinstrument: Kwaliteit van competentie assessment [Evaluation instrument for the quality of competence assessment].* Arnhem, The Netherlands: Cito.